Flexible joint modelling of longitudinal and survival data: The stjm command

17th Stata UK Users' Group Meeting

Michael J. Crowther[1*],
Keith R. Abrams[1] and Paul C. Lambert[1,2]

[1]Centre for Biostatistics and Genetic Epidemiology
Department of Health Sciences
University of Leicester, UK.

[2]Department of Medical Epidemiology and Biostatistics
Karolinska Institutet
Stockholm, Sweden.

*mjc76@le.ac.uk

Outline

- Introduction to joint modelling
- A new joint model
- `stjm`
- Example application - Primary Biliary Cirrhosis (PBC)
- Future work

Background

- Longitudinal response data affected by informative dropout
- Inclusion of time-varying covariates in survival analyses

Background

- Longitudinal response data affected by informative dropout
- Inclusion of time-varying covariates in survival analyses

Background

- Longitudinal response data affected by informative dropout
- Inclusion of time-varying covariates in survival analyses

Approaches:

- Latent class approach (Proust-Lima and Taylor, 2009)
- Shared parameter models - dependence through shared random effects (Wulfsohn and Tsiatis, 1997)

Background

- Longitudinal response data affected by informative dropout
- Inclusion of time-varying covariates in survival analyses

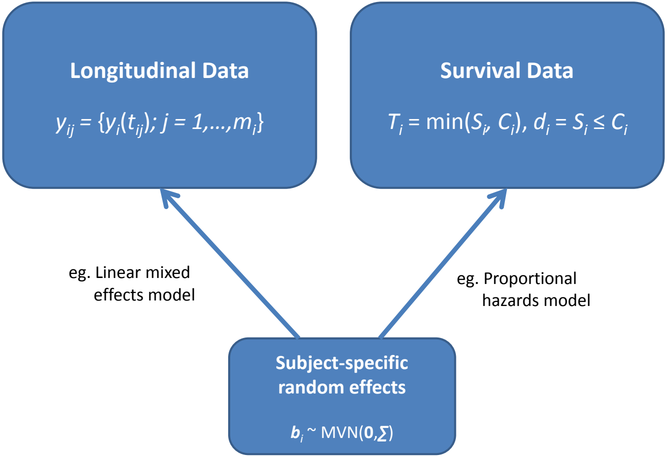
Approaches:

- Latent class approach (Proust-Lima and Taylor, 2009)
- Shared parameter models - dependence through shared random effects (Wulfsohn and Tsiatis, 1997)

Example dataset

- Primary Biliary Cirrhosis (PBC) dataset - 312 patients with 1945 repeated measurements of serum bilirubin (Murtagh et al., 1994).
- 158 randomised to receive D-penicillamine and 154 to placebo
- Interested in the treatment effect after adjusting for the longitudinal biomarker

Data structure



Longitudinal Data

$y_{ij} = \{y_i(t_{ij}); j = 1,\ldots,m_i\}$

Survival Data

$T_i = \min(S_i, C_i), d_i = S_i \leq C_i$

eg. Linear mixed effects model

eg. Proportional hazards model

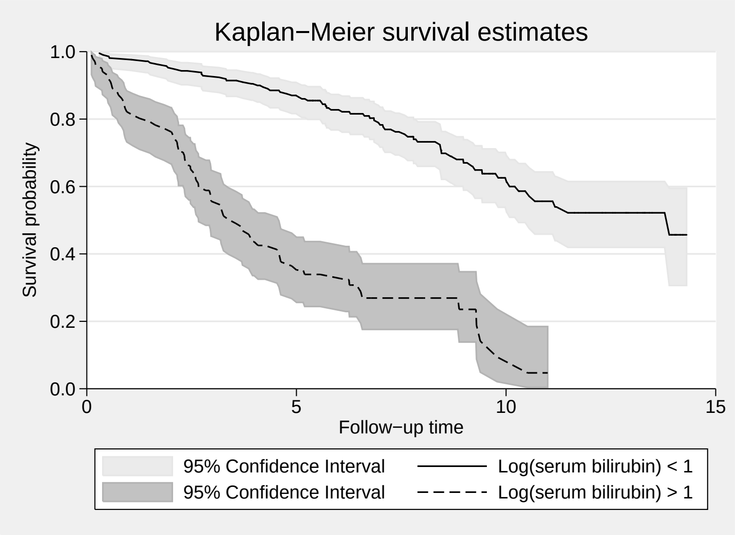
Subject-specific random effects

$\boldsymbol{b}_i \sim \text{MVN}(\mathbf{0},\Sigma)$

Data structure

```
. list id logb drug _t0 _t _d if id==3 | id==20, noobs sepby(id)

  +-------------------------------------------------------------+
  | id       logb        drug        _t0          _t        _d |
  |-------------------------------------------------------------|
  |  3   .3364722   D-penicil          0    .48187494        0 |
  |  3   .0953102   D-penicil  .48187494    .99660498        0 |
  |  3   .4054651   D-penicil  .99660498    2.0342789        0 |
  |  3   .5877866   D-penicil  2.0342789    2.7707808        1 |
  |-------------------------------------------------------------|
  | 20   1.629241     placebo          0    .49556455        0 |
  | 20   2.727853     placebo  .49556455    .91446722        0 |
  | 20   2.406945     placebo  .91446722    3.6797721        0 |
  | 20   3.465736     placebo  3.6797721    3.7126274        1 |
  +-------------------------------------------------------------+
```

Data structure



```
. stjmgraph logb, panel(id)
```

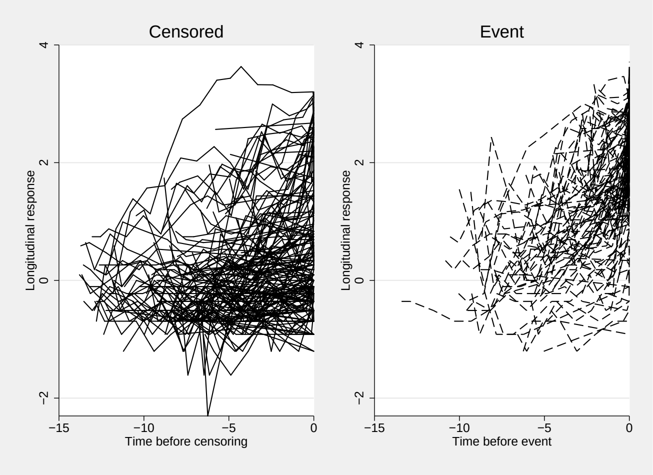


Figure: Longitudinal trajectories. Adjusted timescale.

Longitudinal submodel

- Linear mixed effects model (Laird and Ware, 1982):

$$y_i(t_{ij}) = W_i(t_{ij}) + e_{ij}, \qquad e_{ij} \sim \mathrm{N}(0, \sigma_e^2)$$

$$W_i(t_{ij}) = x_i'(t_{ij})\beta + z_i'(t_{ij})b_i + u_i\delta$$

```
.  xtmixed logb time drug || id:  time, cov(unstr)
```

- Increase flexibility through the use of fixed/random fractional polynomials of time (Royston and Altman, 1994).

Survival submodel

- Many choices proposed including standard parametric and, of course, Cox proportional hazards models

$$h_i(T_i) = h_0(T_i)\exp(\alpha W_i(T_i) + \phi v_i)$$

where v_i is a set of baseline covariates and, for example;

$$W_i(T_i) = \beta_{0i} + \beta_{1i} T_i + \delta u_i$$

- α is termed the association parameter; in this case we assume the association is based on the "current value" of the biomarker

We can now write down the full joint likelihood:

$$\prod_{i=1}^{n} \left[\int_{-\infty}^{\infty} \left(\prod_{j=1}^{m_i} f(y_i(t_{ij})|b_i, \theta) \right) f(b_i|\theta) f(T_i, d_i|b_i, \theta) \; \mathrm{d}b_i \right]$$

where

$$f(y_i(t_{ij})|b_i, \theta) = (2\pi\sigma_e^2)^{-1/2} \text{exp} \left\{ -\frac{y_i(t_{ij}) - W_i(t_{ij})}{2\sigma_e^2} \right\},$$

$$f(b_i|\theta) = (2\pi|V|)^{-1/2} \text{exp} \left\{ -\frac{b_i' V^{-1} b_i}{2} \right\},$$

and

$$f(T_i, d_i|b_i, \theta) = [h_0(T_i) \text{exp}(\alpha W_i(t) + \phi v_i)]^{d_i} \, \text{exp} \left\{ -\int_0^{T_i} h_0(u) \text{exp}(\alpha W_i(u) + \phi v_i) du \right\}$$

We can now write down the full joint likelihood:

$$\prod_{i=1}^{n}\left[\int_{-\infty}^{\infty}\left(\prod_{j=1}^{m_i} f(y_i(t_{ij})|b_i,\theta)\right) f(b_i|\theta) f(T_i,d_i|b_i,\theta)\ \mathrm{d}b_i\right]$$

where

$$f(y_i(t_{ij})|b_i,\theta) = (2\pi\sigma_e^2)^{-1/2}\exp\left\{-\frac{y_i(t_{ij}) - W_i(t_{ij})}{2\sigma_e^2}\right\},$$

$$f(b_i|\theta) = (2\pi|V|)^{-1/2}\exp\left\{-\frac{b_i' V^{-1} b_i}{2}\right\},$$

and

$$f(T_i,d_i|b_i,\theta) = [h_0(T_i)\exp(\alpha W_i(t) + \phi v_i)]^{d_i} \exp\left\{-\int_0^{T_i} h_0(u)\exp(\alpha W_i(u) + \phi v_i)du\right\}$$

Gauss-Hermite quadrature

- Numerical method to approximate analytically intractable integrals (Pinheiro and Bates, 1995)

$$\int_{-\infty}^{\infty} e^{-x^2} f(x)dx \approx \sum_{q=1}^{m} w_q f(x_q)$$

- Can be extended to multivariate integrals i.e. multiple random effects

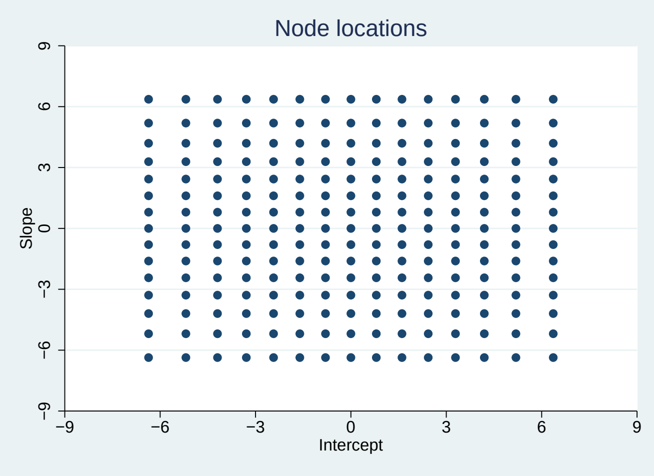


Figure: 2-dimensional 15 point node locations.

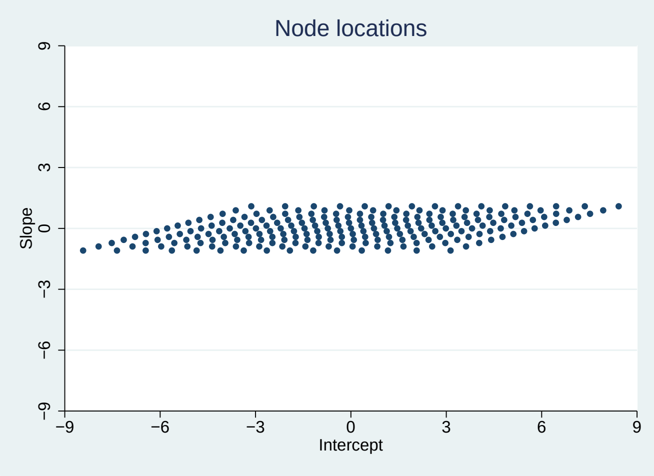


Figure: Adjusted 2-dimensional 15 point node locations.

Survival submodel

- Flexible parametric survival model (Royston and Parmar, 2002; Lambert and Royston, 2009)
- Modelled on the log cumulative hazard scale using restricted cubic splines (Durrleman and Simon, 1989)

$$\log\{H_0(t)\} = s\{\log(t)|\gamma, \mathbf{k}_0\}$$

- Can evaluate the likelihood directly

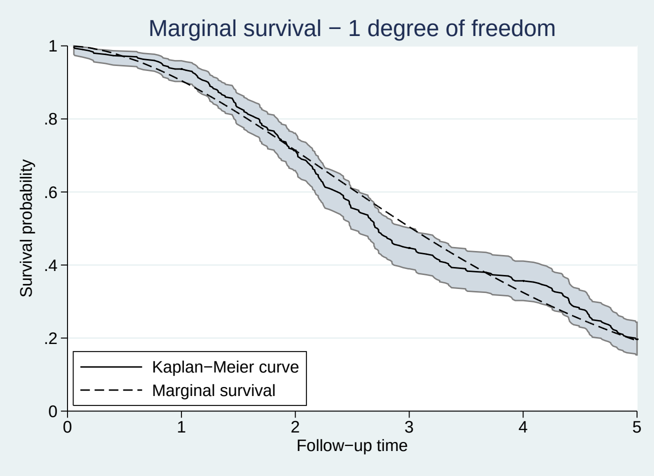


Figure: Predicted marginal survival function from joint model with 1 degree of freedom, overlaid on the Kaplan-Meier survival curve.

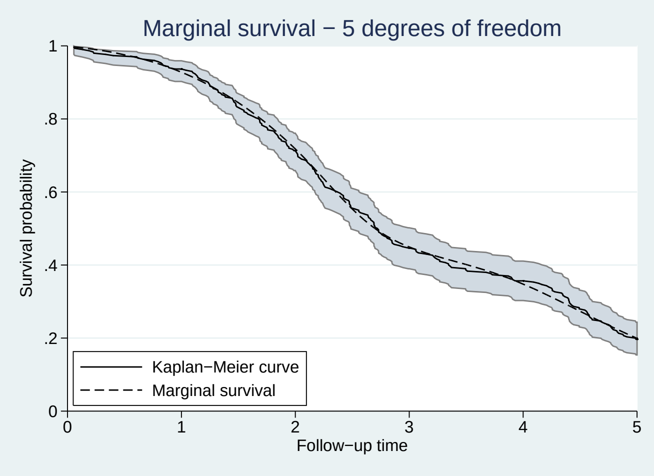


Figure: Predicted marginal survival function from joint model with 5 degrees of freedom, overlaid on the Kaplan-Meier survival curve.

Survival model linear predictor:

$$\log\{H(T_i|\mathbf{b}_i, v_i)\} = \eta_i = s\{\log(T_i)|\gamma, \mathbf{k}_0\} + \alpha W_i(T_i) + \phi v_i$$

Survival model linear predictor:

$$\log\{H(T_i|\mathbf{b}_i, v_i)\} = \eta_i = s\{\log(T_i)|\gamma, \mathbf{k}_0\} + \alpha W_i(T_i) + \phi v_i$$

Transform to the hazard and survival scales:

$$h(T_i|\mathbf{b}_i, v_i) = \left\{ \frac{1}{T_i} \frac{\mathrm{d}s\{\log(T_i)|\gamma, \mathbf{k}_0\}}{\mathrm{d}\log(T_i)} + \alpha \frac{\mathrm{d}W(T_i)}{\mathrm{d}T_i} \right\} \exp(\eta_i)$$

$$S(T_i|\mathbf{b}_i, v_i) = \exp\{-\exp(\eta_i)\}$$

(Crowther et al., 2011)

Implementation in Stata

```
stjm longdepvar [varlist], panel(varname) df(#)
    [nodes(#) ...]
```

Implementation in Stata

`stjm` *longdepvar* [*varlist*], `panel(`*varname*`)` `df(#)` [`nodes(#)` ...]

- Longitudinal submodel:
  - `ffracpoly(`*numlist*`)` - Fixed FP's of time
  - `rfracpoly(`*numlist*`)` - Random FP's of time
  - [*varlist*] - Baseline covariates

Implementation in Stata

`stjm` *longdepvar* [*varlist*], `panel(`*varname*`) df(#)` `[nodes(#) ...]`

- Longitudinal submodel:
  - `ffracpoly(`*numlist*`)` - Fixed FP's of time
  - `rfracpoly(`*numlist*`)` - Random FP's of time
  - [*varlist*] - Baseline covariates
- Survival submodel:
  - `df(#)/knots(`*numlist*`)` - Baseline cum. hazard
  - `survcov(`*varlist*`)` - Baseline covariates

Implementation in Stata

`stjm` *`longdepvar`* `[`*`varlist`*`], panel(`*`varname`*`) df(#) [nodes(#) ...]`

- Longitudinal submodel:
    - `ffracpoly(`*`numlist`*`)` - Fixed FP's of time
    - `rfracpoly(`*`numlist`*`)` - Random FP's of time
    - `[`*`varlist`*`]` - Baseline covariates
- Survival submodel:
    - `df(#)/knots(`*`numlist`*`)` - Baseline cum. hazard
    - `survcov(`*`varlist`*`)` - Baseline covariates
- Association:
    - `nocurrent` - Current value is the default
    - `derivassoc(1)` - 1^{st} derivative
    - `sepintassoc/sepassoc(numlist)` - Random coefficient, e.g. random intercept

Predictions

```
predict newvarname, option
```

Predictions

`predict` *newvarname, option*

- Longitudinal:
  - `xb/fitted` - Fitted values
  - `residuals` - Subject level residuals
  - `rstandard` - Standardised residuals
  - `reffects/reses` - Empirical Bayes predictions of random effects

Predictions

`predict` *newvarname*, *option*

- Longitudinal:
  - `xb/fitted` - Fitted values
  - `residuals` - Subject level residuals
  - `rstandard` - Standardised residuals
  - `reffects/reses` - Empirical Bayes predictions of random effects
- Survival:
  - `hazard` - Hazard function
  - `survival` - Survival function
  - `cumhazard` - Cumulative hazard function
  - `martingale` - Martingale residuals
  - `stjmcondsurv` - Conditional survival

Predictions

`predict` *newvarname*, *option*

- Longitudinal:
    - `xb/fitted` - Fitted values
    - `residuals` - Subject level residuals
    - `rstandard` - Standardised residuals
    - `reffects/reses` - Empirical Bayes predictions of random effects
- Survival:
    - `hazard` - Hazard function
    - `survival` - Survival function
    - `cumhazard` - Cumulative hazard function
    - `martingale` - Martingale residuals
    - `stjmcondsurv` - Conditional survival

Predictions can be evaluated at measurement/survival times, or user specified times.

Predictions

- Random intercept with fixed slope, current value association, one degree of freedom:

  ```
  .  stjm logb drug, ffracp(1) nodes(15) df(1) survcov(drug)
  ```

- Random intercept and slope with fixed time powers 2 and 3, association based on 1st derivative, 3 degrees of freedom:

  ```
  .  stjm logb drug, rfracp(1) ffracp(2 3) survcov(drug) ///
  nodes(15) df(3) nocurrent derivassoc(1)
  ```

Application to PBC dataset

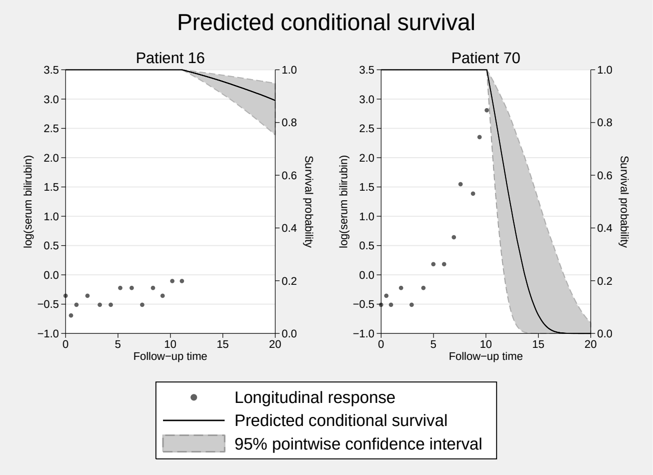
```
. stjm logb trt, panel(id) nodes(15) rfracp(1) df(1) survcov(trt)
-> gen double timevar_1 = X^(1)
(where X = _t0)

Obtaining initial values:

Fitting full model:

Joint model estimates                           Number of obs.      =    1945
Patient variable: id                            Number of patients =     312
Log-likelihood = -1952.7411
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Longitud.:   |
   timevar_1 |   .1636943   .0042372    38.63   0.000     .1553895     .171999
         trt |  -.1705512   .0349558    -4.88   0.000    -.2390633   -.1020391
       _cons |    .682979   .0348529    19.60   0.000     .6146686    .7512894
-------------+----------------------------------------------------------------
Survival:    |
         trt |  -.0209648    .177859    -0.12   0.906     -.369562    .3276325
       _rcs1 |   .8489682   .0820932    10.34   0.000     .6880684    1.009868
       _cons |  -3.330624   .2450019   -13.59   0.000    -3.810819   -2.850429
-------------+----------------------------------------------------------------
Association: |
     current |   1.010613   .0836087    12.09   0.000     .8467429    1.174483
------------------------------------------------------------------------------
```

Random effects table

```
------------------------------------------------------------------------------
  Random effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
id: Unstructured             |
               sd(timevar_1) |   .1588024   .0055511      .1482869    .1700637
               sd(intercept) |   .8564879   .0173008      .8232413    .8910771
   corr(timevar_1,intercept) |   .5405518   .0208844      .4983468     .580201
-----------------------------+------------------------------------------------
                sd(Residual) |   .3687027   .0067102      .3557827    .3820918
------------------------------------------------------------------------------
 Longitudinal submodel: Linear mixed effects model
     Survival submodel: Flexible parametric model with 1 degree of freedom
    Integration method: Gauss-Hermite quadrature using 15 quadrature points
```

Individual level predictions

```
stjmcondsurv, panel(id) id(16) fu(20)
```



Future work

- Survival submodels:
  - Weibull PH model
  - Gompertz PH model
  - 2-component mixture Weibull PH model
  - Mixture Weibull-exponential PH model
- Extension to competing risks
- Extension to include a cure proportion
- Longitudinal categorical responses
- EM algorithm
- Adaptive GH quadrature

Command acknowlegdments

- `rcsgen` - Paul Lambert
- `stpm2` - Paul Lambert
- `ghquadm` - Bill Sribney
- `esttab` - Ben Jann
- `fracgen` - Patrick Royston

References I

M. J. Crowther, K. R. Abrams, and P. C Lambert. Flexible parametric joint modelling of longitudinal and survival data. *Submitted*, 2011.

S. Durrleman and R. Simon. Flexible regression models with cubic splines. *Statistics in Medicine*, 8(5):551–561, 1989.

N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974, 1982.

P. C Lambert and P. Royston. Further development of flexible parametric models for survival analysis. *The Stata Journal*, 9:265–290, 2009.

P. Murtagh, E. Dickson, M. Van Dam, G. Malincho, and P. Grambsch. Primary biliary cirrhosis: Prediction of short-term survival based on repeated patient visits. *Hepatology*, 20:126–134, 1994.

José C. Pinheiro and Douglas M. Bates. Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, 4(1):pp. 12–35, 1995. ISSN 10618600.

C. Proust-Lima and J. M. G. Taylor. Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment psa: a joint modeling approach. *Biostatistics*, 10(3): 535–549, 2009.

P. Royston and D. G. Altman. Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 43(3):429–467, 1994.

P. Royston and M. K. B. Parmar. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Stat Med*, 21(15):2175–2197, 2002.

M. S. Wulfsohn and A. A. Tsiatis. A joint model for survival and longitudinal data measured with error. *Biometrics*, 53(1):330–339, 1997.