Flexible parametric joint modelling of longitudinal and survival data

Workshop on Flexible Models for Longitudinal and Survival Data with Applications in Biostatistics
University of Warwick
27th - 29th July 2015

Michael J. Crowther[1,2,*]

[1] *Department of Health Sciences*
University of Leicester, UK
[2] *Department of Medical Epidemiology and Biostatistics*
Karolinska Institutet, Sweden
[*] michael.crowther@le.ac.uk

Outline

Background

stjm

Multivariate JMs

Delayed entry

JMs in large datasets

Summary

Outline

Background

stjm

Multivariate JMs

Delayed entry

JMs in large datasets

Summary

Background

- Biomarkers, such as blood pressure, are often collected repeatedly over time, in parallel to the time to an event of interest, such as death from any cause

Background

- Biomarkers, such as blood pressure, are often collected repeatedly over time, in parallel to the time to an event of interest, such as death from any cause
- These biomarkers are often measured with error

Background

- Biomarkers, such as blood pressure, are often collected repeatedly over time, in parallel to the time to an event of interest, such as death from any cause
- These biomarkers are often measured with error
- Issues:
  - Longitudinal studies are often affected by (informative) drop-out, e.g. due to death
  - Can we account for measurement error when looking at how a time-varying biomarker is associated with an event of interest?

How can we link longitudinal and survival data?

How can we link longitudinal and survival data?

- Use the observed baseline biomarker values
  - We're ignoring all the repeated measures and measurement error

How can we link longitudinal and survival data?

- Use the observed baseline biomarker values
  - We're ignoring all the repeated measures and measurement error
- Use the repeated measures as a time-varying covariate
  - We're still ignoring the measurement error

How can we link longitudinal and survival data?

- Use the observed baseline biomarker values
  - We're ignoring all the repeated measures and measurement error
- Use the repeated measures as a time-varying covariate
  - We're still ignoring the measurement error
- Model the longitudinal outcome, and use predictions as a time-varying covariate
  - Uncertainty in the longitudinal outcome is not carried through

How can we link longitudinal and survival data?

- Use the observed baseline biomarker values
  - We're ignoring all the repeated measures and measurement error
- Use the repeated measures as a time-varying covariate
  - We're still ignoring the measurement error
- Model the longitudinal outcome, and use predictions as a time-varying covariate
  - Uncertainty in the longitudinal outcome is not carried through
- Model both processes simultaneously in a joint model
  - Reduce bias and maximise efficiency

Joint modelling of longitudinal and survival data

- Arose primarily in the field of AIDS, relating CD4 trajectories to progression to AIDS in HIV positive patients (Faucett and Thomas, 1996)
- Further developed in cancer, particularly modelling PSA levels and their association with prostate cancer recurrence (Proust-Lima and Taylor, 2009)

Joint modelling of longitudinal and survival data

- Arose primarily in the field of AIDS, relating CD4 trajectories to progression to AIDS in HIV positive patients (Faucett and Thomas, 1996)
- Further developed in cancer, particularly modelling PSA levels and their association with prostate cancer recurrence (Proust-Lima and Taylor, 2009)

Two core methodological approaches have arisen

- Latent class approach (Proust-Lima et al., 2012)
- Shared parameter models - dependence through shared random effects (Wulfsohn and Tsiatis, 1997; Henderson et al., 2000; Gould et al., 2014)

Joint modelling of longitudinal and survival data

- Arose primarily in the field of AIDS, relating CD4 trajectories to progression to AIDS in HIV positive patients (Faucett and Thomas, 1996)
- Further developed in cancer, particularly modelling PSA levels and their association with prostate cancer recurrence (Proust-Lima and Taylor, 2009)

Two core methodological approaches have arisen

- Latent class approach (Proust-Lima et al., 2012)
- Shared parameter models - dependence through shared random effects (Wulfsohn and Tsiatis, 1997; Henderson et al., 2000; Gould et al., 2014)

The basic framework

Longitudinal submodel

Assume we observe continuous longitudinal marker:

$$y_i(t) = m_i(t) + e_i(t), \qquad e_i(t) \sim \mathrm{N}(0, \sigma^2)$$

where

$$m_i(t) = \boldsymbol{X}_i^T(t)\boldsymbol{\beta} + \boldsymbol{Z}_i^T(t)\boldsymbol{b}_i + \boldsymbol{u}_i^T\boldsymbol{\delta}$$

and

$$\boldsymbol{b_i} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$$

Flexibility can be incorporated through fractional polynomials or splines in X_i and Z_i.

The basic framework

Survival submodel

We assume a proportional hazards survival submodel

$$h_i(t) = h_0(t) \exp\left[\boldsymbol{\phi}^T \boldsymbol{v}_i + \alpha m_i(t)\right]$$

where $h_0(t)$ is the baseline hazard function, and $\boldsymbol{v}_i \in \boldsymbol{U}_i$ a set of baseline time-independent covariates with associated vector of log hazard ratios, $\boldsymbol{\phi}$.

The basic framework

Survival submodel

We assume a proportional hazards survival submodel

$$h_i(t) = h_0(t) \exp\left[\boldsymbol{\phi}^T \boldsymbol{v}_i + \alpha m_i(t)\right]$$

where $h_0(t)$ is the baseline hazard function, and $\boldsymbol{v}_i \in \boldsymbol{U}_i$ a set of baseline time-independent covariates with associated vector of log hazard ratios, $\boldsymbol{\phi}$.

Linking the component models

Our key question here is how are changes in the biomarker trajectory associated with survival?

$$h_i(t) = h_0(t) \exp\left[\boldsymbol{\phi}^T \boldsymbol{v}_i + \alpha m_i(t)\right]$$

where for example

$$m_i(t) = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t + \boldsymbol{u}_i^T \boldsymbol{\delta}$$

$\alpha m_i(t)$ is termed the current value parameterisation

Linking the component models

Our key question here is how are changes in the biomarker trajectory associated with survival?

$$h_i(t) = h_0(t) \exp\left[\boldsymbol{\phi}^T \boldsymbol{v}_i + \alpha m_i(t)\right]$$

where for example

$$m_i(t) = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t + \boldsymbol{u}_i^T \boldsymbol{\delta}$$

$\alpha m_i(t)$ is termed the current value parameterisation

Alternative association structures

Interaction effects

$$h_i(t) = h_0(t) \exp\left[\phi^T \boldsymbol{v_{i1}} + \alpha^T \{\boldsymbol{v_{i2}} \times m_i(t)\}\right]$$

where $\boldsymbol{v_{i1}}, \boldsymbol{v_{i2}} \in \boldsymbol{U_i}$. We now have vector of association parameters α, providing different associations for different covariate patterns.

Alternative association structures

Time-dependent slope

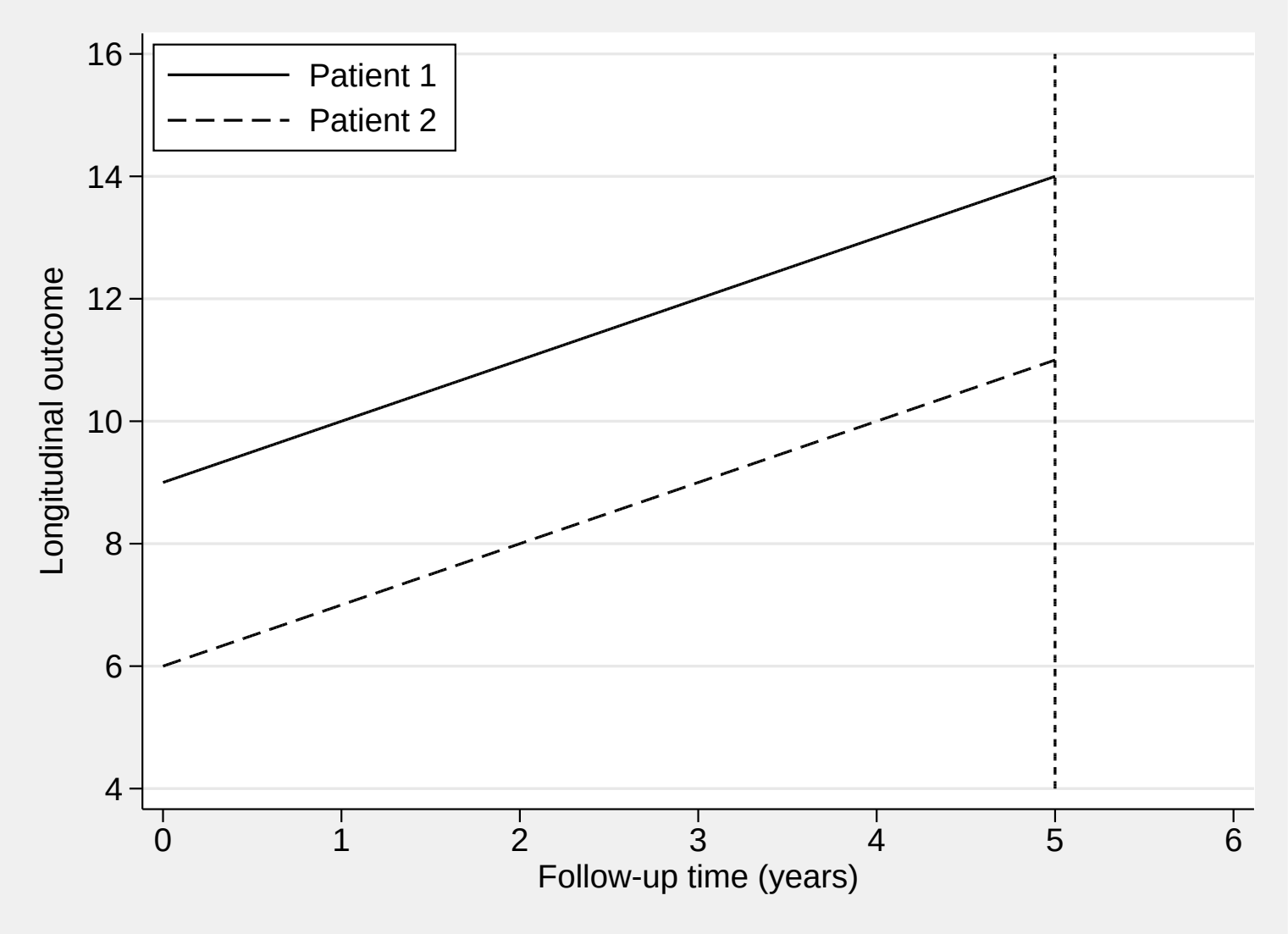
We may be interested in how the slope or rate of change of the biomarker is associated with survival:

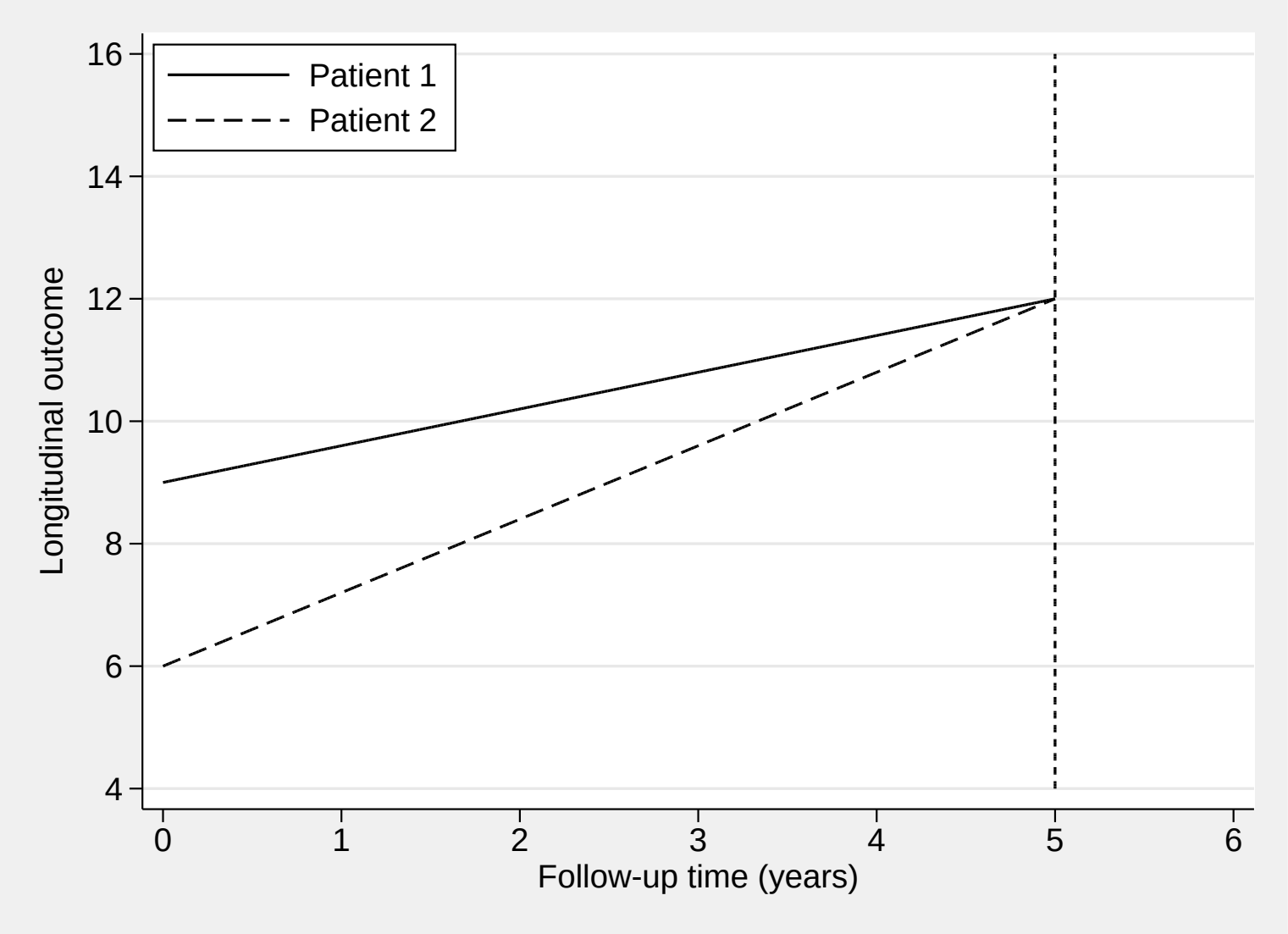
$$h_i(t) = h_0(t) \exp\left[\boldsymbol{\phi}^T \boldsymbol{v_i} + \alpha_1 m_i(t) + \alpha_2 m_i'(t)\right] \quad (1)$$

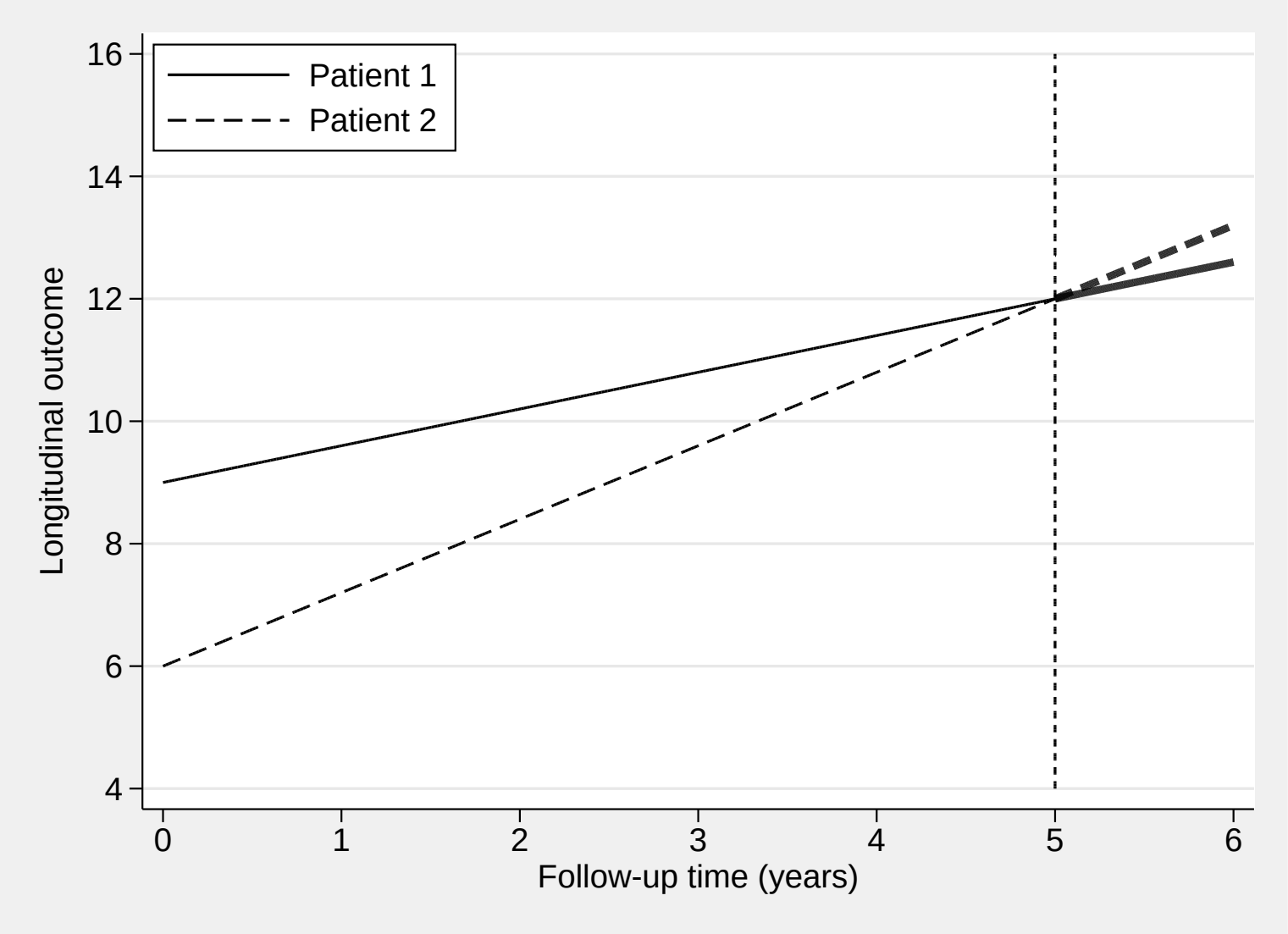
with

$$m_i'(t) = \frac{\mathrm{d}m_i(t)}{\mathrm{d}t} = \frac{\mathrm{d}\{\boldsymbol{X}_i^T(t)\boldsymbol{\beta} + \boldsymbol{Z}_i^T(t)\boldsymbol{b_i}\}}{\mathrm{d}t} \quad (2)$$

The added benefit of including the rate of change of CD4 trajectories within a joint model framework to model the risk of progression to AIDS or death in HIV-positive patients was conducted by Wolbers et al. (2010).


Patient 1
Patient 2
Longitudinal outcome
16
14
12
10
8
6
4
0
1
2
3
4
5
6
Follow-up time (years)


Patient 1
Patient 2
Longitudinal outcome
16
14
12
10
8
6
4
0
1
2
3
4
5
6
Follow-up time (years)


Patient 1
Patient 2
Longitudinal outcome
16
14
12
10
8
6
4
0
1
2
3
4
5
6
Follow-up time (years)

Alternative association structures

Random effects parameterisation

Finally, I define a time-independent association structure

$$h_i(t) = h_0(t) \exp\left[\boldsymbol{\phi}^T \boldsymbol{v}_i + \boldsymbol{\alpha}^T(\boldsymbol{\beta} + \boldsymbol{b}_i)\right] \quad (3)$$

Equation (3) includes both the population level mean of the random effect, plus the subject specific deviation, for example

$$h_i(t) = h_0(t) \exp\left[\boldsymbol{\phi}^T \boldsymbol{v}_i + \alpha_1(\beta_0 + b_{0i})\right] \quad (4)$$

where $\exp(\alpha_1)$ is the hazard ratio for a one unit increase in the baseline value of the longitudinal outcome i.e. the intercept (Crowther et al., 2013b).

Joint likelihood

The full joint likelihood is

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

Joint likelihood

The full joint likelihood is

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

where

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

Joint likelihood

The full joint likelihood is

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

where

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

Joint likelihood

The full joint likelihood is

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

where

$$p(T_i, d_i|b_i,\theta) = [h_0(T_i)\exp(\alpha m_i(t) + \phi v_i)]^{d_i}$$

$$\times\exp\left\{-\int_0^{T_i} h_0(u)\exp(\alpha m_i(u) + \phi v_i)du\right\}$$

Joint likelihood

The full joint likelihood is

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

where

$$p(T_i,d_i|b_i,\theta) = [h_0(T_i)\exp(\alpha m_i(t)+\phi v_i)]^{d_i}$$

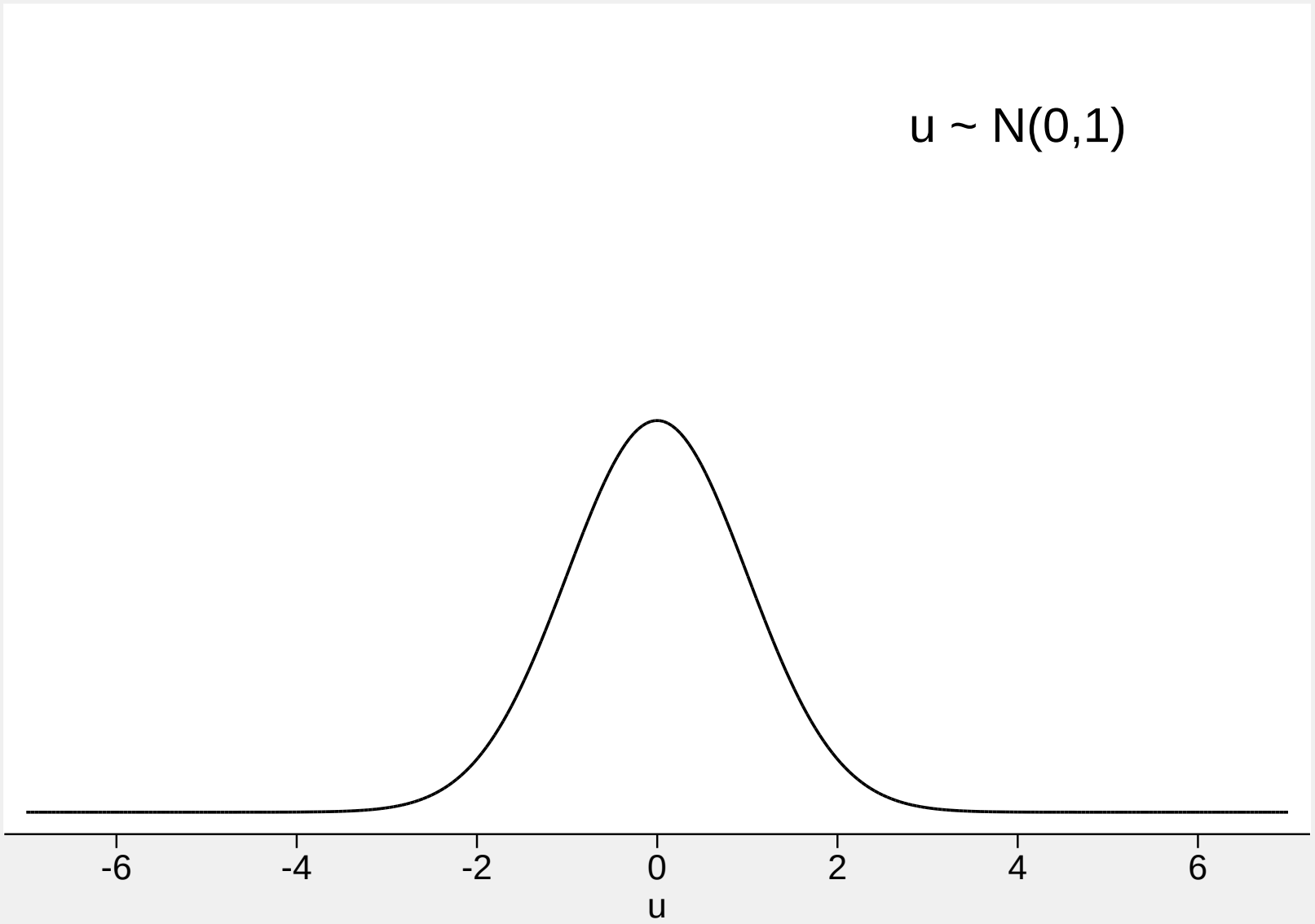
$$\times\exp\left\{-\int_0^{T_i} h_0(u)\exp(\alpha m_i(u)+\phi v_i)du\right\}$$

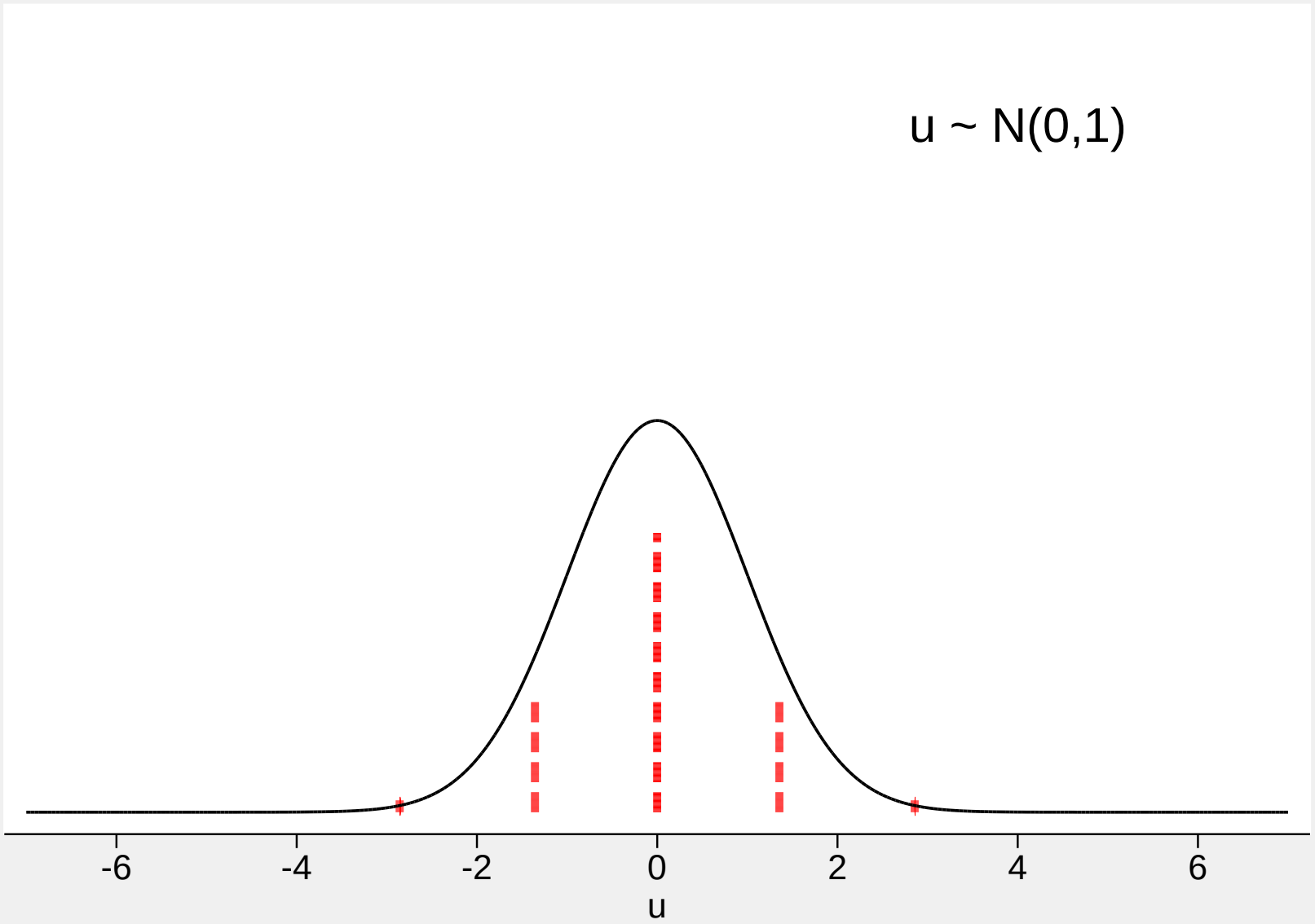
Gauss-Hermite quadrature

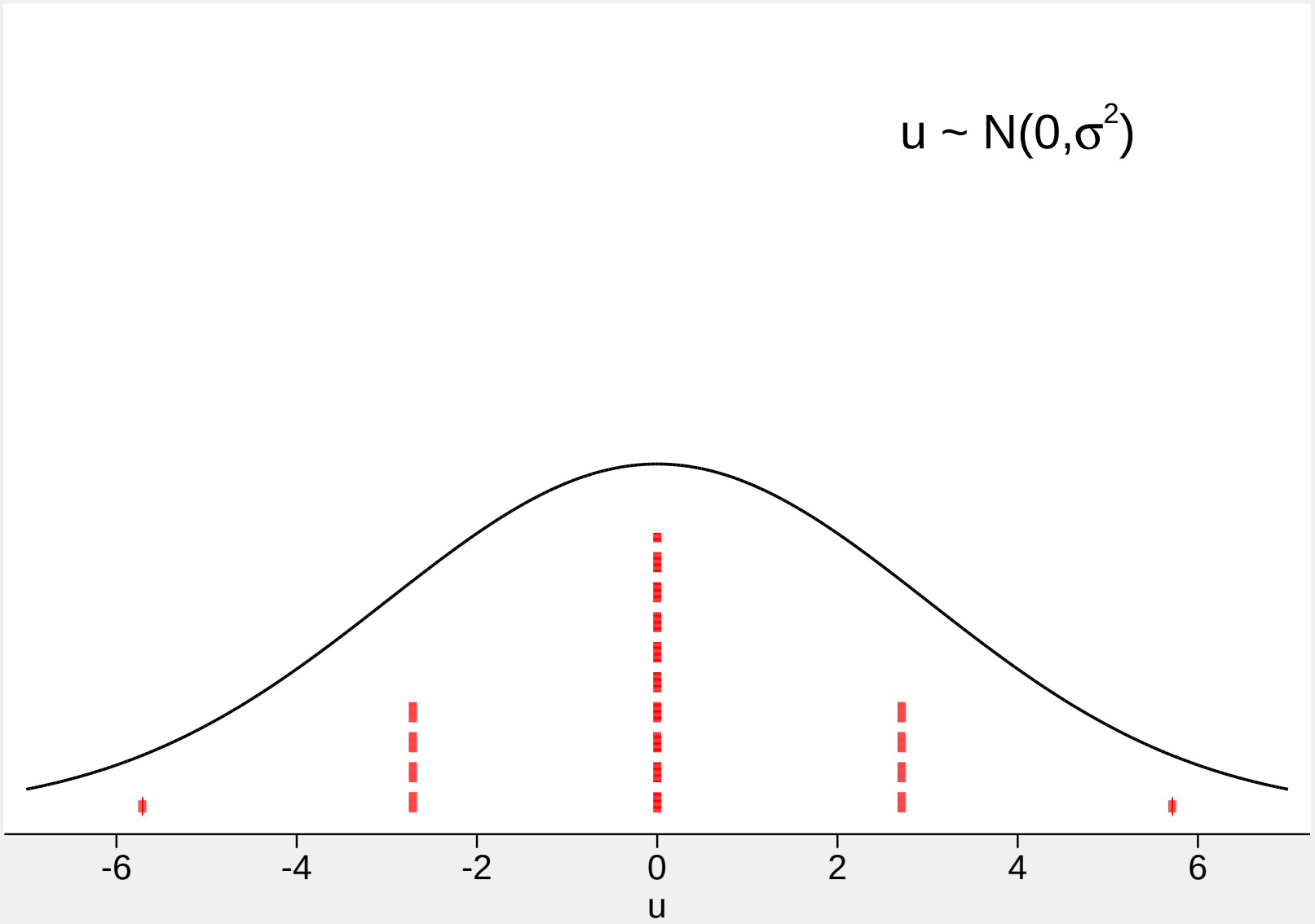
- Numerical method to approximate analytically intractable integrals (Pinheiro and Bates, 1995)

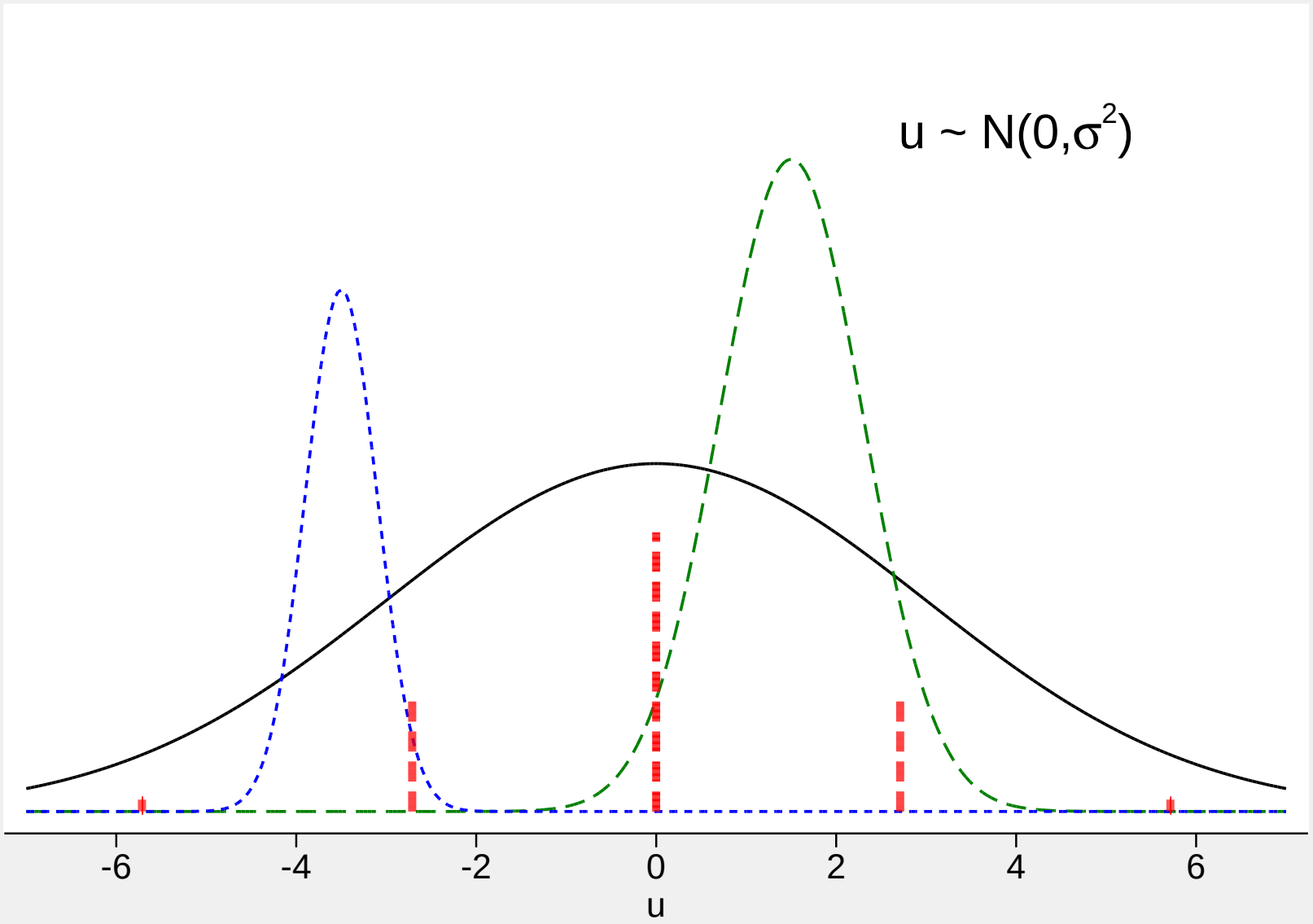
$$\int_{-\infty}^{\infty} e^{-x^2} f(x) dx \approx \sum_{q=1}^{m} w_q f(x_q)$$

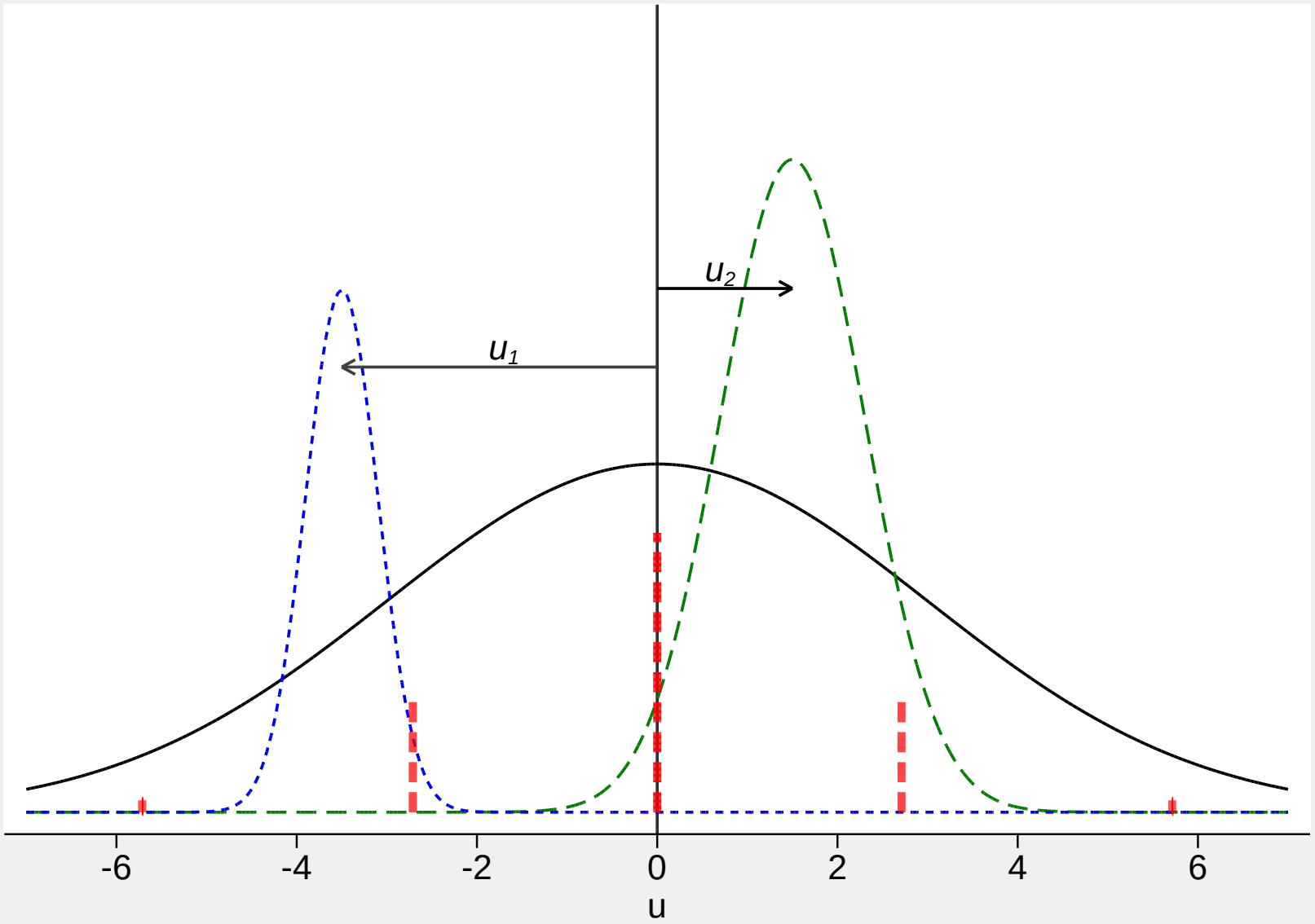
- Can be extended to multivariate integrals i.e. multiple random effects

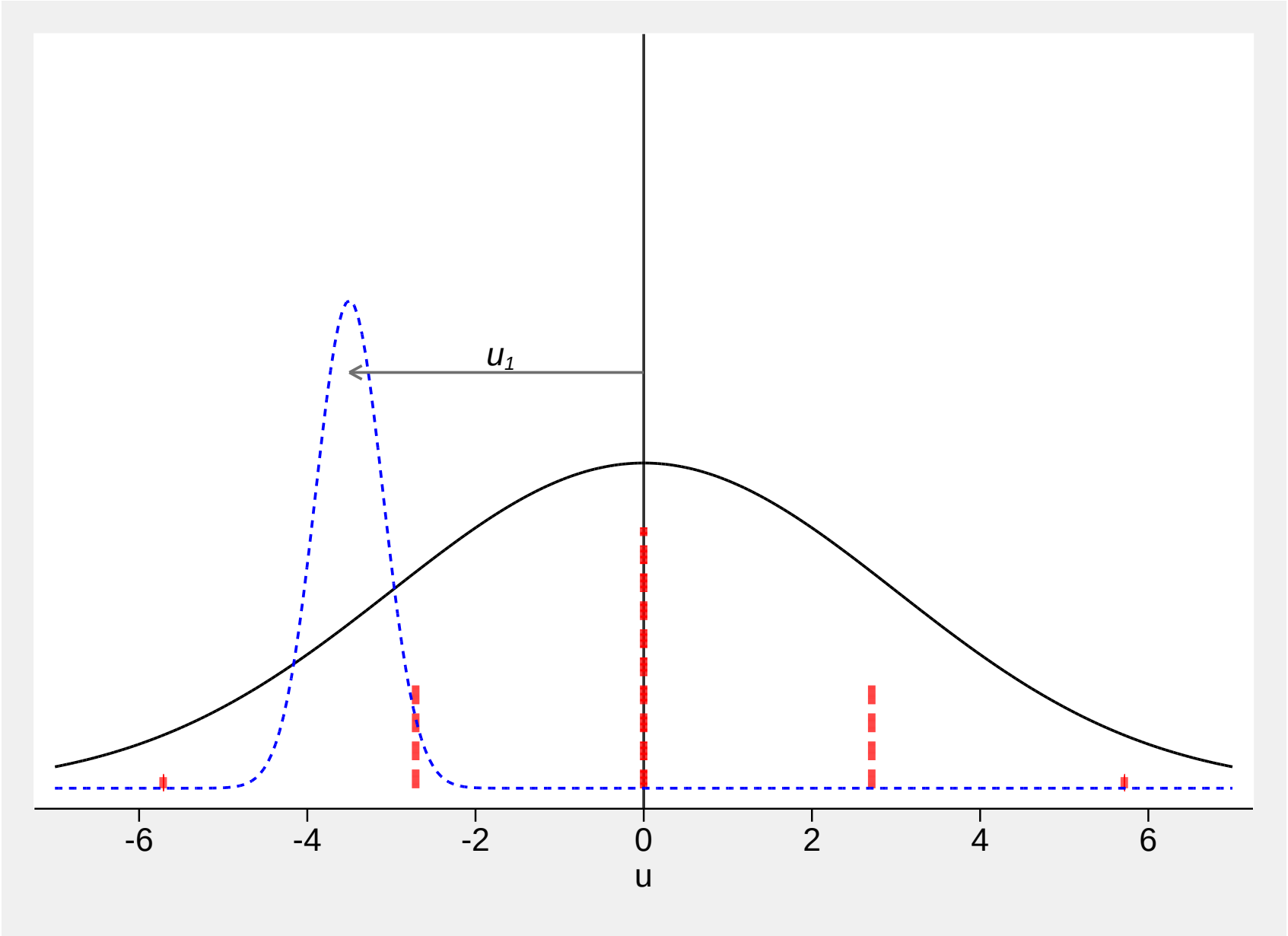

u ~ N(0,1)
-6
-4
-2
0
2
4
6
u

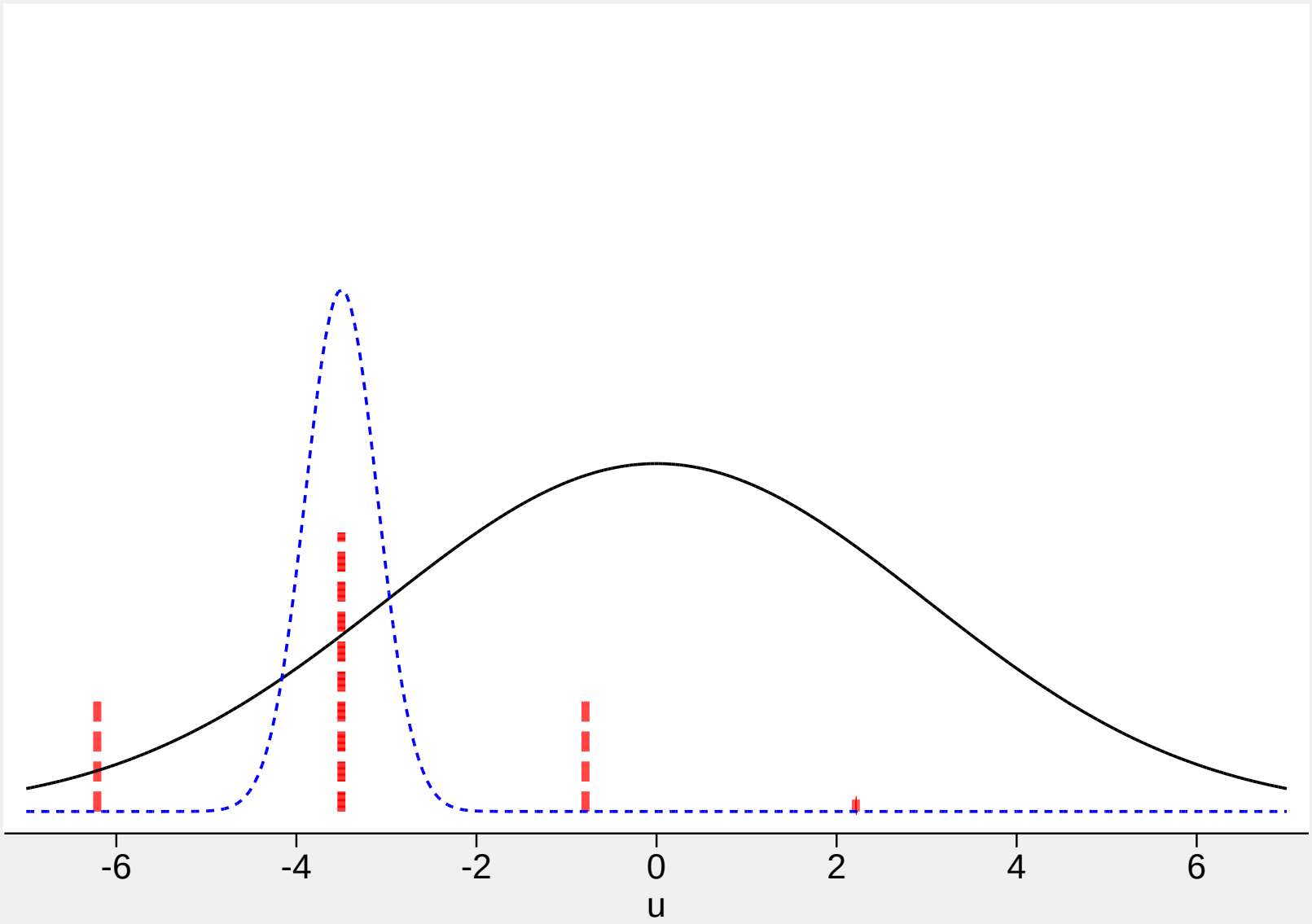

u ~ N(0,1)
-6
-4
-2
0
2
4
6
u

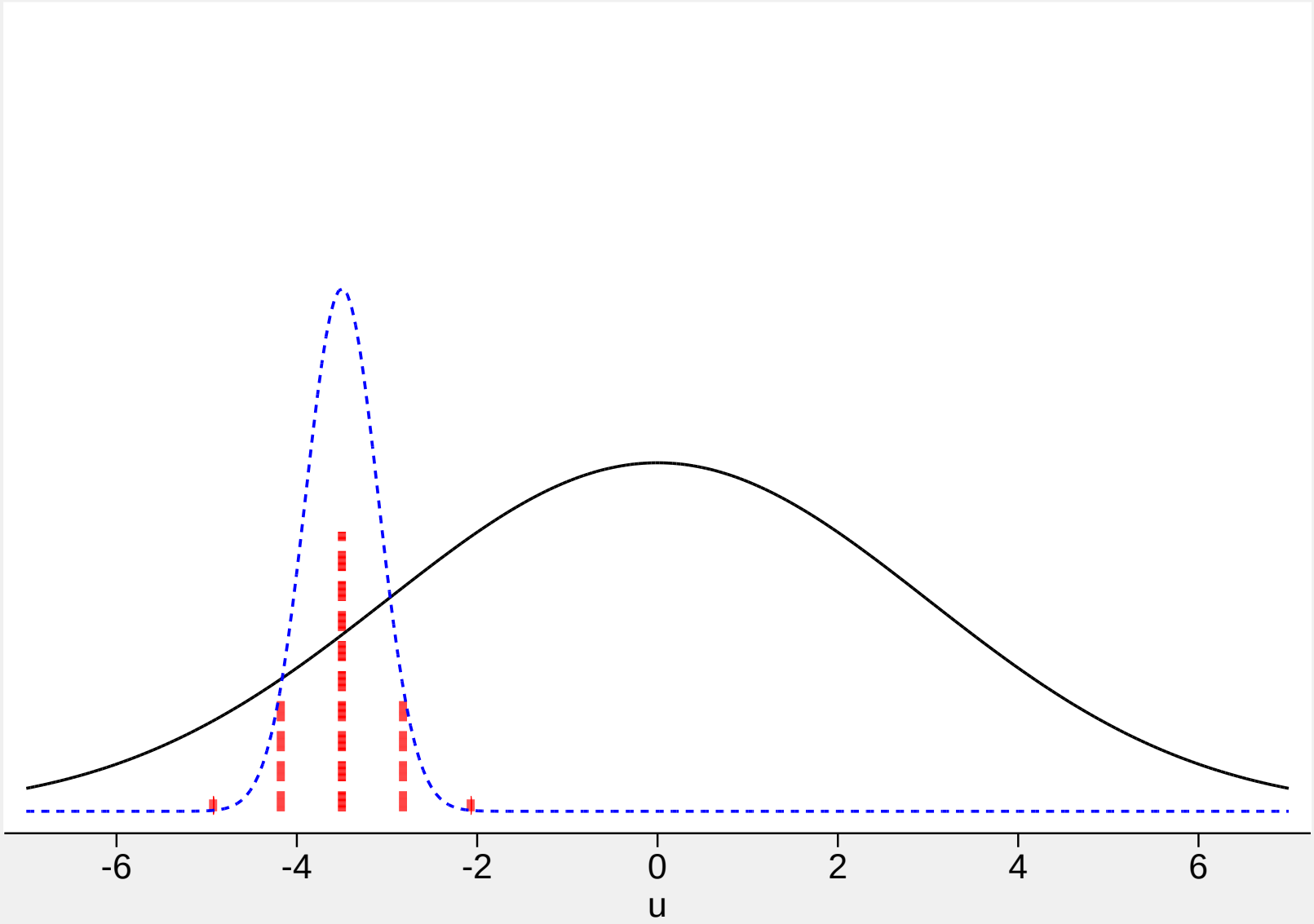

u ~ N(0,σ²)
-6
-4
-2
0
2
4
6
u


u ~ N(0,σ²)
-6
-4
-2
0
2
4
6
u


u_2
u_1
-6
-4
-2
0
2
4
6
u


u_1
-6
-4
-2
0
2
4
6
u


-6
-4
-2
0
2
4
6
u


-6
-4
-2
0
2
4
6
u

Outline

Background

stjm

Multivariate JMs

Delayed entry

JMs in large datasets

Summary

D-penicillamine for primary biliary cirrhosis - RCT

- Data from 312 patients with PBC collected at the Mayo Clinic 1974-1984 (Murtaugh et al., 1994)
- 158 randomised to receive D-penicillamine and 154 to placebo
- Outcome is all-cause death with 140 events observed
- 1945 measurements of serum bilirubin
- Interested in the association between serum bilirubin over time and survival

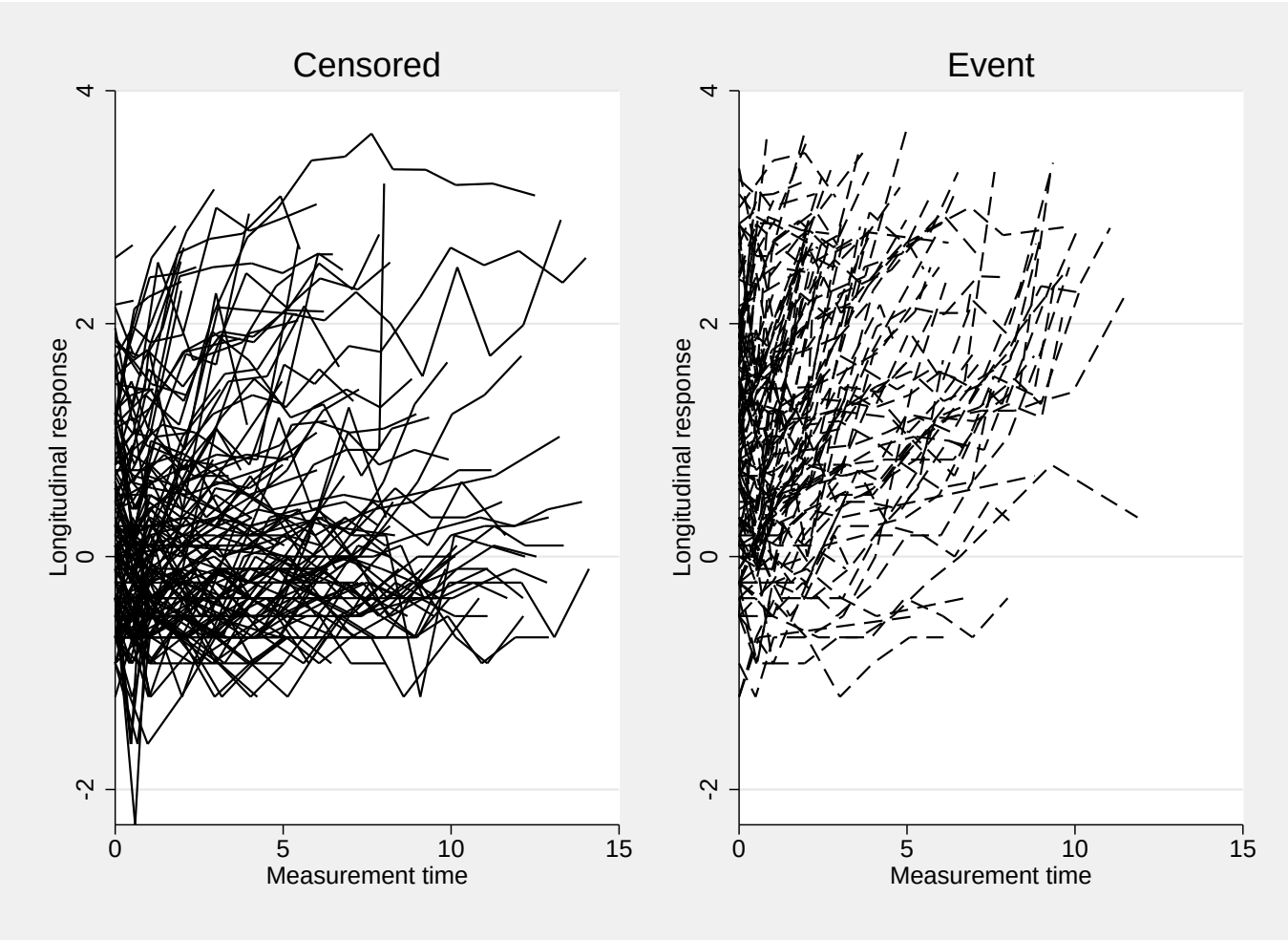
Data structure - survival data in Stata

```
. stset stop, enter(start) failure(died=1) id(id)
```

id	logb	trt	_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
48	.6418539	0	0	.52294379	0
48	.1823216	0	.52294379	1.0431497	0
48	.7884574	0	1.0431497	1.9658307	0
48	.5306283	0	1.9658307	2.9569597	0
48	.4054651	0	2.9569597	3.9782062	0
48	.2623642	0	3.9782062	4.9885006	0
48	.3364722	0	4.9885006	5.9495125	0
48	.1823216	0	5.9495125	6.8913593	0
48	.3364722	0	6.8913593	13.95247	0

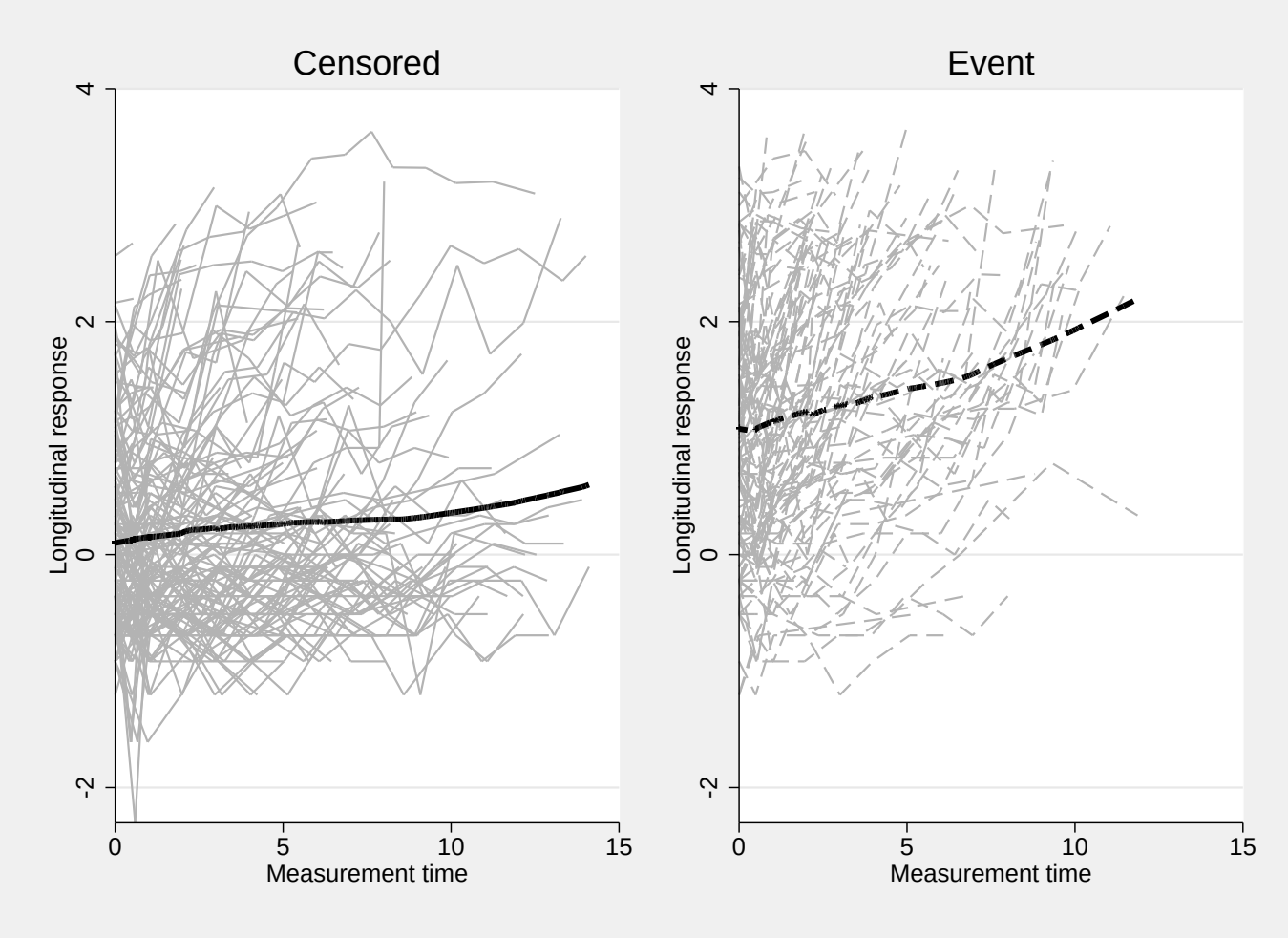
.

Exploratory trajectory plots



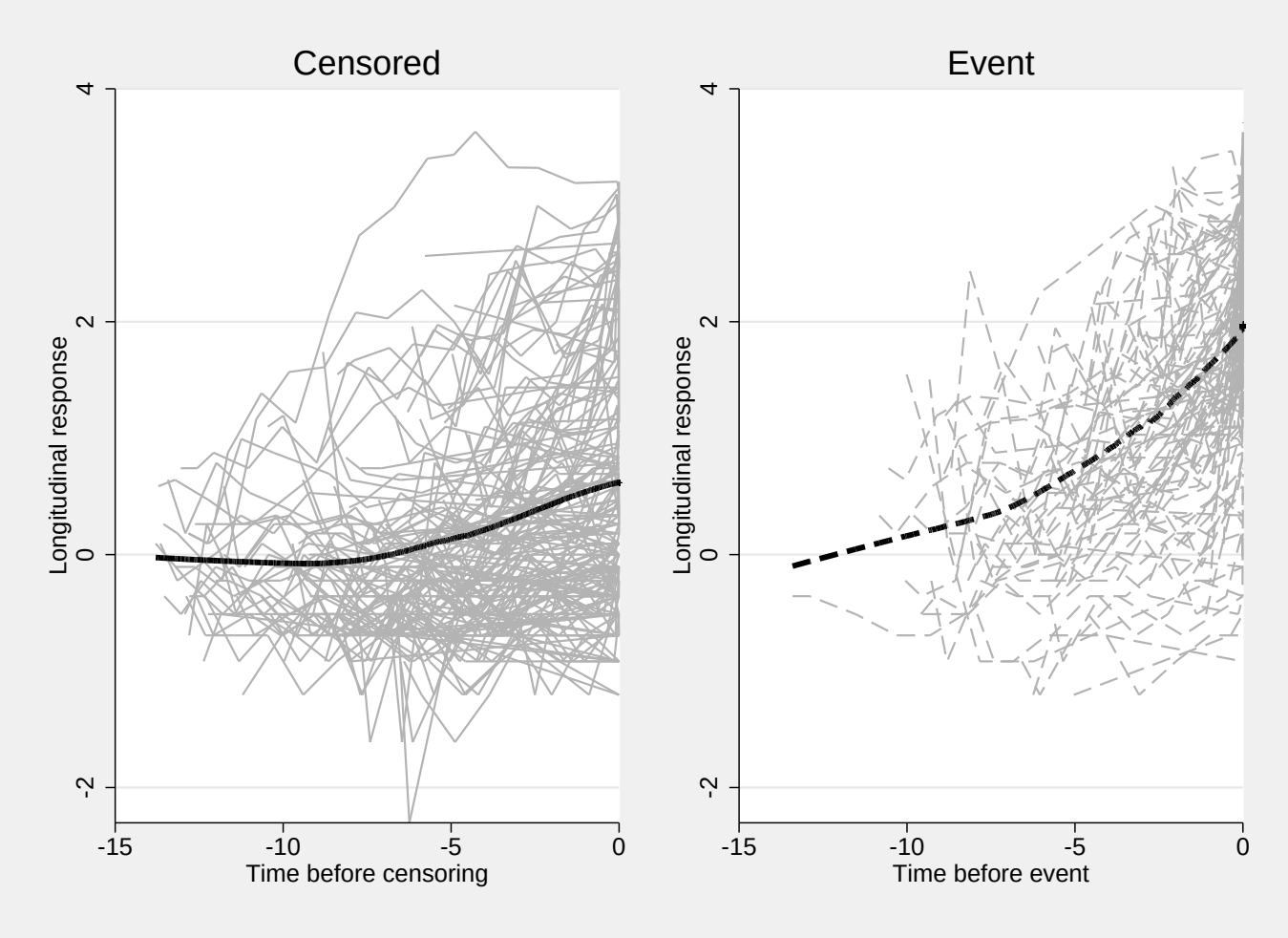

```
. stjmgraph logb, panel(id)
```

Exploratory trajectory plots




```
. stjmgraph prothrombin, panel(id) lowess
```

Exploratory trajectory plots




```
stjmgraph prothrombin, panel(id) lowess adjust
```

Syntax

```
. stjm longdepvar [varlist],
>                                    panel(varname)
>                                    survmodel(model)
>                                    [options]
```

Syntax

```
. stjm longdepvar [varlist],
>                                    panel(varname)
>                                    survmodel(model)
>                                    [options]
```

- Longitudinal submodel:
  - `[varlist]` - Baseline covariates
  - `ffp(numlist)` - Fixed FP's of time
  - `rfp(numlist)` - Random FP's of time
  - `frcs(#)` - Fixed RCS of time
  - `rrcs(#)` - Random RCS of time
  - `timeinteraction(varlist)` - covariates to interact with fixed time variables
  - `covariance(vartype)` - variance-covariance structure of random effects

Syntax

```
. stjm longdepvar [varlist],
>                                    panel(varname)
>                                    survmodel(model)
>                                    [options]
```

- Survival submodel:
  - `survmodel(`*model*`)` - Survival model including exponential, Weibull, Gompertz, splines on the log hazard scale, Royston-Parmar, 2-component mixtures
  - `survcov(`*varlist*`)` - Baseline covariates
  - `df(#)` - degrees of freedom for baseline hazard function
  - `knots(`*numlist*`)` - internal knot locations
  - `noorthog` - suppress default orthogonalisation

Syntax

```
. stjm longdepvar [varlist],
>                                    panel(varname)
>                                    survmodel(model)
>                                    [options]
```

- Association:
  - `nocurrent` - Current value is the default
  - `derivassoc` - 1^{st} derivative (slope)
  - `intassoc`/`assoc(numlist)` - Random coefficient, e.g. random intercept
  - `nocoefficient` - do not include fixed coefficient in time-independent associations
  - `assoccov(`*varlist*`)` adjust the association parameter(s) by covariates

Syntax

```
. stjm longdepvar [varlist],
>                                    panel(varname)
>                                    survmodel(model)
>                                    [options]
```

- Maximisation:
  - `gh(#)` - number of Gauss-Hermite quadrature nodes
  - `gk(#)` - number of Gauss-Kronrod quadrature nodes
  - `adaptit(#)` - number of adaptive quadrature iterations; default is 5
  - `nonadapt` - use non-adaptive Gauss-Hermite quadrature
  - `showadapt` - display iteration log for adaptive sub-routine

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 can be evaluated at measurement/survival times, or user specified times (`timevar(`*`varname`*`)`), and at specific covariate patterns using `at(`*`varname # ...`*`)`.

Predictions

`predict` *`newvarname, option`*

- 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 (`timevar(`*`varname`*`)`), and at specific covariate patterns using `at(`*`varname # ...`*`)`.

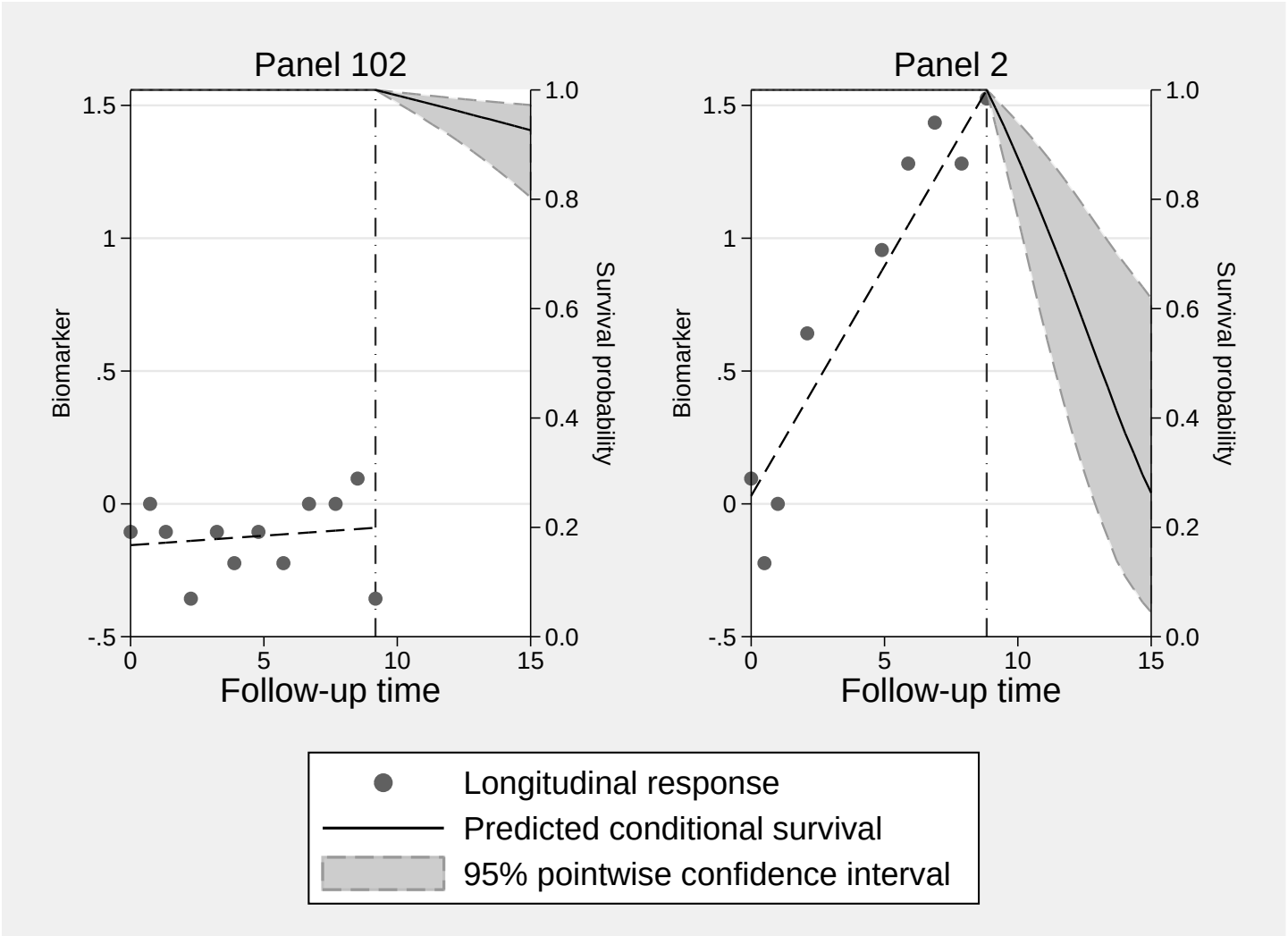
```
. stjm logb, panel(id) survmodel(weibull) rfp(1) timeinterac(trt) survcov(trt)
-> gen double _time_1 = X^(1)
-> gen double _time_1_trt = trt * _time_1
(where X = _t0)
Obtaining initial values:
Fitting full model:
-> Conducting adaptive Gauss-Hermite quadrature
Iteration 0:   log likelihood = -1923.9358
Iteration 1:   log likelihood = -1919.2078
Iteration 2:   log likelihood = -1919.1856
Iteration 3:   log likelihood = -1919.1855
Joint model estimates                             Number of obs      =      1945
Panel variable: id                                Number of panels   =       312
                                                  Number of failures =       140
Log-likelihood = -1919.1855
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Longitudinal						
_time_1	.1826073	.0183264	9.96	0.000	.1466883	.2185264
_time_1_trt	.0045744	.0244713	0.19	0.852	-.0433885	.0525373
_cons	.4927346	.0582861	8.45	0.000	.3784959	.6069733
Survival						
assoc:value						
_cons	1.24083	.0931223	13.32	0.000	1.058314	1.423347
ln_lambda						
trt	.0407589	.179847	0.23	0.821	-.3117347	.3932525
_cons	-4.409684	.2740596	-16.09	0.000	-4.946831	-3.872537
ln_gamma						
_cons	.0188928	.0827694	0.23	0.819	-.1433322	.1811178

Random effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
id: Unstructured				
sd(_time_1)	.1806879	.0123806	.1579812	.2066583
sd(_cons)	1.002541	.0426659	.9223098	1.089751
corr(_time_1,_cons)	.4256451	.0728762	.2730232	.5573664
sd(Residual)	.3471453	.0066734	.334309	.3604745

```
 Longitudinal submodel: Linear mixed effects model
     Survival submodel: Weibull proportional hazards model
    Integration method: Adaptive Gauss-Hermite quadrature using 5 nodes
     Cumulative hazard: Gauss-Kronrod quadrature using 15 nodes
.
```

```
. stjmcondsurv, panel(id) id(102)
```



Computation time

- Dataset of 312 patients with 1945 observations, random intercept and slope model with current value parameterisation computation time is **17 seconds**
- Registry based data example of 5,000 patients with $\approx$ 100,000 measurements takes $\approx$ **10 minutes**

Outline

Background

stjm

Multivariate JMs

Delayed entry

JMs in large datasets

Summary

Multivariate generalised linear mixed joint longitudinal-survival model

- In many cases, we may have multiple (possibly correlated) longitudinal outcomes

Multivariate generalised linear mixed joint longitudinal-survival model

- In many cases, we may have multiple (possibly correlated) longitudinal outcomes
  - Cardiovascular outcomes - blood pressure, serum cholesterol, BMI
  - HIV - CD4 cell counts, viral load
  - Primary biliary cirrhosis example - serum bilirubin, prothrombin index

Multivariate generalised linear mixed joint longitudinal-survival model

- In many cases, we may have multiple (possibly correlated) longitudinal outcomes
  - Cardiovascular outcomes - blood pressure, serum cholesterol, BMI
  - HIV - CD4 cell counts, viral load
  - Primary biliary cirrhosis example - serum bilirubin, prothrombin index
- They may not all be continuous outcomes

Multivariate generalised linear mixed joint longitudinal-survival model

We have $k = 1, \ldots, K$ longitudinal responses, therefore we let $\boldsymbol{y}_i(t) = (y_{i1}(t), \ldots, y_{iK}(t))^T$ denote a vector of K longitudinal observations for the i^{th} patient at time t.

$$g_k(m_{ik}(t)) = \boldsymbol{X}_{ik}^T(t)\boldsymbol{\beta}_k + \boldsymbol{Z}_{ik}^T(t)\boldsymbol{b}_{ik} + \boldsymbol{u}_{ik}^T\boldsymbol{\delta}_k$$

where $g_k(.)$ is a known link function.

$$\begin{pmatrix} \boldsymbol{b}_{i1} \\ \vdots \\ \boldsymbol{b}_{iK} \end{pmatrix} = \boldsymbol{b}_i \sim \mathrm{N}(0, \boldsymbol{V})$$

Multivariate generalised linear mixed joint longitudinal-survival model

The survival submodel follows

$$h(t|\boldsymbol{M}_i(t), \boldsymbol{v_i}) = h_0(t) \exp\left(\boldsymbol{\phi}^T \boldsymbol{v_i} + \sum_{k=1}^{K} \alpha_k m_{ik}(t)\right)$$

Likelihood

We define the log-likelihood, assuming conditional independence, and censoring is non-informative

$$l(\theta) = \sum_i \log \int_{\boldsymbol{b}_i} \left[\prod_{k=1}^{K} \prod_{j=1}^{n_i} p(y_{ikj}|\boldsymbol{\theta}) \right] p(T_i, d_i|\boldsymbol{\theta}) p(\boldsymbol{b}_i|\boldsymbol{\theta}) d\boldsymbol{b}_i$$

where the j^{th} observation for the i^{th} patient of the k^{th} longitudinal outcome is y_{ikj}. We assume the longitudinal component comes from the exponential family, of the form

$$p(y_{ikj}|\boldsymbol{\theta}) = \exp(B_k[\mu_{ik}(t_{ij})]y_{ik}(t_{ij}) - A_k[\mu_{ik}(t_{ij})] + C_k[y_{ik}(t_{ij})])$$

with $A_k[.]$, $B_k[.]$, and $C_k[.]$, appropriate functions for the type of longitudinal outcome in question.

Returning to the PBC example

- We actually have multiple markers available, including prothrombin index, albumin levels,...
- We can fit a multivariate joint model, with serum bilirubin and prothrombin index as continuous longitudinal outcomes
- For simplicity, I'll assume a random intercept and fixed linear slope for each trajectory, and allow the random intercepts to be correlated
- I assume the current value association for both outcomes

```
. stjm logb, ffp(1) || pro , panel(id) ffp(1) survmodel(weibull) survcov(trt)
-> gen double _time_1_1 = X^(1)
(where X = _t0)
-> gen double _time_2_1 = X^(1)
(where X = _t0)
Obtaining initial values:
Fitting full model:
-> Conducting adaptive Gauss-Hermite quadrature
Iteration 0:   log likelihood = -5677.2212  (not concave)
Iteration 1:   log likelihood =  -5623.156
Iteration 2:   log likelihood =  -5609.665
Iteration 3:   log likelihood = -5608.4292
Iteration 4:   log likelihood = -5608.4081
Iteration 5:   log likelihood = -5608.4081
Joint model estimates                             Number of obs      =      1945
Panel variable: id                                Number of panels   =       312
                                                  Number of failures =       140

Log-likelihood = -5608.4081
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Longitudin~1						
_time_1_1	.0980146	.0043117	22.73	0.000	.0895638	.1064654
_cons1	.5801024	.0650802	8.91	0.000	.4525476	.7076572
Longitudin~2						
_time_2_1	.1492314	.0103142	14.47	0.000	.1290159	.1694469
_cons2	10.73592	.0634756	169.13	0.000	10.61151	10.86033

Survival						
A1:value						
_cons	.9350057	.1368665	6.83	0.000	.6667522	1.203259
A2:value						
_cons	.6880163	.1821244	3.78	0.000	.3310589	1.044974
ln_lambda						
trt	.0102147	.1804639	0.06	0.955	-.3434881	.3639175
_cons	-11.87136	1.998969	-5.94	0.000	-15.78927	-7.953455
ln_gamma						
_cons	.1003182	.0788456	1.27	0.203	-.0542162	.2548527

Random effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
id: Unstructured				
sd(_cons1)	1.109695	.0471279	1.021066	1.206018
sd(_cons2)	.8525335	.0525647	.7554903	.9620419
corr(_cons1,_cons2)	.732477	.0402468	.6433418	.8020195
sd(Residual1)	.4913156	.0085897	.4747652	.508443
sd(Residual2)	1.243429	.0218243	1.201381	1.286948

Longitudinal submodel: Linear mixed effects model
Survival submodel: Weibull proportional hazards model
Integration method: Adaptive Gauss-Hermite quadrature using 5 nodes
Cumulative hazard: Gauss-Kronrod quadrature using 15 nodes

- Each longitudinal outcome can be linked to survival in many different ways

- Each longitudinal outcome can be linked to survival in many different ways
- Each longitudinal outcome can be modelled flexibly using splines or polynomials

- Each longitudinal outcome can be linked to survival in many different ways
- Each longitudinal outcome can be modelled flexibly using splines or polynomials
- We can specify a variance-covariance structure, and also impose constraints on the variance-covariance matrix of the random effects

- Each longitudinal outcome can be linked to survival in many different ways
- Each longitudinal outcome can be modelled flexibly using splines or polynomials
- We can specify a variance-covariance structure, and also impose constraints on the variance-covariance matrix of the random effects
- We can use the `family(gauss|binomial|poisson)` option for each of the outcomes

Outline

Background

stjm

Multivariate JMs

Delayed entry

JMs in large datasets

Summary

Delayed entry

- Predominantly joint modelling has been applied to clinical trial data
- Moving to applications using registry data raises a number of methodological issues
  - Defining a clinically meaningful time origin can often be difficult
  - From an epidemiological perspective, age is often used as the timescale, as it can be an improved way of controlling for the effect of age (Thiébaut and Bénichou, 2004)
  - This then requires delayed entry (left truncation)

Likelihood with delayed entry

$$L_i(.|T_i > T_{0i}) = \frac{\int \left[\prod_{j=1}^{n_i} p(y_{ij}|\boldsymbol{b}_i, \boldsymbol{\theta})\right] p(T_i, d_i|\boldsymbol{b}_i, \boldsymbol{\theta}) p(\boldsymbol{b}_i|\boldsymbol{\theta}) d\boldsymbol{b}_i}{S(T_{0i}|\boldsymbol{\theta})}$$

where

$$S(T_{0i}|\boldsymbol{\theta}) = \int S(T_{0i}|\boldsymbol{b}_i, \boldsymbol{\theta}) p(\boldsymbol{b}_i|\boldsymbol{\theta}) d\boldsymbol{b}_i$$

Both the numerator and denominator require numerical integration

Joint models with delayed entry:

- Proust-Lima et al. (2009) within a joint latent class model
- Dantan et al. (2011) developed a joint model for longitudinal data and an illness-death process in cognitive aging. Estimated using non-adaptive Gauss-Hermite quadrature
- van den Hout and Muniz-Terrera (2015) developed a joint model for repeatedly measured discrete data, arising from tests of cognitive function, and survival in the analysis of an older population. Assumed a random intercept and linear slope in the longitudinal submodel, assessing a binomial or beta-binomial formulation, and a Weibull or Gompertz survival model. As interest was in prediction, they emphasised the benefits of adopting a parametric survival submodel. Non-adaptive quadrature was used.

Mammographic breast density and survival in breast cancer

- Re-analysis of a study in breast cancer (Li et al., 2013) which looked at mammographic breast density reduction between first and second mammograms
- Longitudinal measurements of mammographic breast density and time to death
- Main hypothesis is that mammographic density reduction after diagnosis among women treated with tamoxifen is a prognostic marker
- We use information on breast cancer from an observational, population-based case-control study conducted in Sweden 1993-1995, who enter the cohort at diagnosis

- The cohort is restricted to patients who have at least 2 mammograms
- By definition this means they cannot die until the time of second mammogram, which therefore requires delayed entry
- Through a joint model we can utilise all available measurements
- Furthermore, we still wish to use the baseline observation in the longitudinal submodel
- We therefore have 974 patients with 6158 images, where 121 (12.4%) patients die before study end

Simulated example

	id	_t0	_t	_d	long_r~e	entry
1.	1	1	1.2	0	.93	1
2.	1	1.2	1.7	0	1.32	1
3.	2	.4	.5	0	1.15	1.3
4.	2	.5	1.2	0	1.67	1.3
5.	2	1.2	1.6	0	1.92	1.3
6.	2	1.6	1.9	0	2.65	1.3
7.	2	1.9	2.6	1	3.15	1.3
8.	3	0	2	0	.25	2
9.	3	2	2.3	0	.21	2
10.	3	2.3	2.4	1	.31	2

```
. stjm long_response, panel(id) rfp(1) survmodel(weibull) enter(enter)
```

Parameter	Linear					Polynomial				
	Estimate	Std. Err.	p-value	95% CI		Estimate	Std. Err.	p-value	95% CI	
Longitudinal										
Time	-0.088	0.004	0.000	-0.096	-0.080	-0.351	0.029	0.000	-0.409	-0.294
Time^2	-	-	-	-	-	0.084	0.013	0.000	0.058	0.109
Time^3	-	-	-	-	-	-0.010	0.002	0.000	-0.014	-0.006
Time^4	-	-	-	-	-	0.000	0.000	0.000	0.000	0.001
Age (years)	-0.029	0.006	0.000	-0.039	-0.018	-0.029	0.006	0.000	-0.040	-0.018
BMI (kg/m^2)	-0.110	0.009	0.000	-0.129	-0.092	-0.111	0.009	0.000	-0.129	-0.093
HR therapy	0.003	0.071	0.972	-0.136	0.141	0.002	0.070	0.975	-0.136	0.140
Tumour size (mm)	0.006	0.004	0.107	-0.001	0.013	0.005	0.004	0.128	-0.002	0.012
Intercept	8.699	0.412	0.000	7.892	9.506	8.925	0.415	0.000	8.112	9.737
Survival										
Association	4.756	4.518	0.293	-4.100	13.612	0.192	0.607	0.752	-0.998	1.382
Age (years)	-0.024	0.017	0.159	-0.058	0.010	-0.017	0.015	0.264	-0.046	0.013
BMI (kg/m^2)	0.013	0.024	0.605	-0.035	0.060	0.016	0.024	0.507	-0.031	0.062
Tamoxifen	0.416	0.206	0.043	0.013	0.818	0.407	0.205	0.047	0.006	0.808
ER status	-0.416	0.259	0.108	-0.922	0.091	-0.458	0.252	0.069	-0.953	0.036
Missing ER status	-0.382	0.313	0.223	-0.996	0.232	-0.443	0.304	0.145	-1.039	0.153
Tumour size (mm)	0.026	0.008	0.001	0.011	0.041	0.024	0.007	0.001	0.010	0.039
No. of metastatic nodes	0.087	0.014	0.000	0.059	0.116	0.088	0.013	0.000	0.062	0.115
Grade = 2	0.610	0.446	0.172	-0.265	1.485	0.617	0.445	0.166	-0.256	1.489
Grade = 3	0.490	0.453	0.279	-0.398	1.377	0.516	0.451	0.253	-0.368	1.399
Grade = Missing	0.616	0.447	0.168	-0.260	1.492	0.647	0.445	0.146	-0.225	1.520
Chemotherapy	0.238	0.318	0.454	-0.385	0.861	0.231	0.312	0.460	-0.381	0.842
Spline 1	-0.028	0.136	0.836	-0.294	0.238	-0.064	0.156	0.682	-0.371	0.242
Spline 2	0.177	0.127	0.162	-0.071	0.426	0.214	0.171	0.212	-0.122	0.549
Intercept	-4.035	1.414	0.004	-6.806	-1.264	-4.910	1.126	0.000	-7.117	-2.703

Parameter	Linear					Polynomial				
	Estimate	Std. Err.	p-value	95% CI		Estimate	Std. Err.	p-value	95% CI	
Longitudinal										
Time	-0.088	0.004	0.000	-0.096	-0.080	-0.351	0.029	0.000	-0.409	-0.294
Time^2	-	-	-	-	-	0.084	0.013	0.000	0.058	0.109
Time^3	-	-	-	-	-	-0.010	0.002	0.000	-0.014	-0.006
Time^4	-	-	-	-	-	0.000	0.000	0.000	0.000	0.001
Age (years)	-0.029	0.006	0.000	-0.039	-0.018	-0.029	0.006	0.000	-0.040	-0.018
BMI (kg/m^2)	-0.110	0.009	0.000	-0.129	-0.092	-0.111	0.009	0.000	-0.129	-0.093
HR therapy	0.003	0.071	0.972	-0.136	0.141	0.002	0.070	0.975	-0.136	0.140
Tumour size (mm)	0.006	0.004	0.107	-0.001	0.013	0.005	0.004	0.128	-0.002	0.012
Intercept	8.699	0.412	0.000	7.892	9.506	8.925	0.415	0.000	8.112	9.737
Survival										
Association	4.756	4.518	0.293	-4.100	13.612	0.192	0.607	0.752	-0.998	1.382
Age (years)	-0.024	0.017	0.159	-0.058	0.010	-0.017	0.015	0.264	-0.046	0.013
BMI (kg/m^2)	0.013	0.024	0.605	-0.035	0.060	0.016	0.024	0.507	-0.031	0.062
Tamoxifen	0.416	0.206	0.043	0.013	0.818	0.407	0.205	0.047	0.006	0.808
ER status	-0.416	0.259	0.108	-0.922	0.091	-0.458	0.252	0.069	-0.953	0.036
Missing ER status	-0.382	0.313	0.223	-0.996	0.232	-0.443	0.304	0.145	-1.039	0.153
Tumour size (mm)	0.026	0.008	0.001	0.011	0.041	0.024	0.007	0.001	0.010	0.039
No. of metastatic nodes	0.087	0.014	0.000	0.059	0.116	0.088	0.013	0.000	0.062	0.115
Grade = 2	0.610	0.446	0.172	-0.265	1.485	0.617	0.445	0.166	-0.256	1.489
Grade = 3	0.490	0.453	0.279	-0.398	1.377	0.516	0.451	0.253	-0.368	1.399
Grade = Missing	0.616	0.447	0.168	-0.260	1.492	0.647	0.445	0.146	-0.225	1.520
Chemotherapy	0.238	0.318	0.454	-0.385	0.861	0.231	0.312	0.460	-0.381	0.842
Spline 1	-0.028	0.136	0.836	-0.294	0.238	-0.064	0.156	0.682	-0.371	0.242
Spline 2	0.177	0.127	0.162	-0.071	0.426	0.214	0.171	0.212	-0.122	0.549
Intercept	-4.035	1.414	0.004	-6.806	-1.264	-4.910	1.126	0.000	-7.117	-2.703

- More in Crowther et al. (2015):
  - Adaptive quadrature
  - Longitudinal trajectory misspecification
- In the model fitted, we linked mammographic density to survival through the current rate of change
- We're actually interested in rate of change soon after start of treatment

Outline

Background

stjm

Multivariate JMs

Delayed entry

JMs in large datasets

Summary

Use of joint models in practice

- Joint models are generally considered to be computationally intensive
- This is predominantly due to the use of numerical integration required to calculate the likelihood
- In general, I think this is overstated; however, in large datasets this becomes fact

I'm working on two large registry based datasets:

I'm working on two large registry based datasets:

1. CALIBER based at the Farr Institute at UCL
   - consists of research ready quantitative data from linked electronic health records and administrative health data. These include CPRD (primary care), HES, national registry data from the Myocardial Ischaemia National Audit Program (MINAP) and mortality and social deprivation data from the ONS
   - 2 million patients, ten million person-years of follow-up

I'm working on two large registry based datasets:

1. CALIBER based at the Farr Institute at UCL
   - consists of research ready quantitative data from linked electronic health records and administrative health data. These include CPRD (primary care), HES, national registry data from the Myocardial Ischaemia National Audit Program (MINAP) and mortality and social deprivation data from the ONS
   - 2 million patients, ten million person-years of follow-up
2. SCANDAT based at Karolinska Institutet in Stockholm
   - quantitative data including repeat haemoglobin levels, linked to cancer diagnoses
   - >1.7 million donors with >25 million donation records

I'm working on two large registry based datasets:

1. CALIBER based at the Farr Institute at UCL
   - consists of research ready quantitative data from linked electronic health records and administrative health data. These include CPRD (primary care), HES, national registry data from the Myocardial Ischaemia National Audit Program (MINAP) and mortality and social deprivation data from the ONS
   - 2 million patients, ten million person-years of follow-up
2. SCANDAT based at Karolinska Institutet in Stockholm
   - quantitative data including repeat haemoglobin levels, linked to cancer diagnoses
   - >1.7 million donors with >25 million donation records

I want to fit joint models...wish me luck!

Joint models in large datasets

- Two-stage models
- Case-cohort methodology

Likelihood with sampling weights

The full joint likelihood is

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

- Sample controls and up-weight
- Use robust standard errors

Outline

Background

stjm

Multivariate JMs

Delayed entry

JMs in large datasets

Summary

Summary

- A wealth of patient data is becoming available in registry sources, as electronic healthcare record linkage moves to the forefront of life science strategy (Jutte et al., 2011)

Summary

- A wealth of patient data is becoming available in registry sources, as electronic healthcare record linkage moves to the forefront of life science strategy (Jutte et al., 2011)
- Current cardiovascular risk scores use observed baseline biomarkers

Summary

- A wealth of patient data is becoming available in registry sources, as electronic healthcare record linkage moves to the forefront of life science strategy (Jutte et al., 2011)
- Current cardiovascular risk scores use observed baseline biomarkers
- Opportunities to utilise the joint model framework in prognostic modelling are great

Summary

- A wealth of patient data is becoming available in registry sources, as electronic healthcare record linkage moves to the forefront of life science strategy (Jutte et al., 2011)
- Current cardiovascular risk scores use observed baseline biomarkers
- Opportunities to utilise the joint model framework in prognostic modelling are great
- Applications so far have been to datasets < 2000 patients

Summary

- A wealth of patient data is becoming available in registry sources, as electronic healthcare record linkage moves to the forefront of life science strategy (Jutte et al., 2011)
- Current cardiovascular risk scores use observed baseline biomarkers
- Opportunities to utilise the joint model framework in prognostic modelling are great
- Applications so far have been to datasets < 2000 patients
- Sensitivity analysis for both processes should be conducted

Summary

- A wealth of patient data is becoming available in registry sources, as electronic healthcare record linkage moves to the forefront of life science strategy (Jutte et al., 2011)
- Current cardiovascular risk scores use observed baseline biomarkers
- Opportunities to utilise the joint model framework in prognostic modelling are great
- Applications so far have been to datasets < 2000 patients
- Sensitivity analysis for both processes should be conducted
- `stjm` will be updated in the next few weeks to include multivariate GLMMs and delayed entry
- `ssc install stjm` (Crowther et al., 2013a)

Featured Article

Statistics in Medicine

Received 27 January 2014, Accepted 19 February 2014 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6141

Joint modeling of survival and longitudinal non-survival data: current methods and issues. Report of the DIA Bayesian joint modeling working group

A. Lawrence Gould,[a*†] Mark Ernest Boye,[b] Michael J. Crowther,[c] Joseph G. Ibrahim,[d] George Quartey,[e] Sandrine Micallef[f] and Frederic Y. Bois[g,h]

Explicitly modeling underlying relationships between a survival endpoint and processes that generate longitudinal measured or reported outcomes potentially could improve the efficiency of clinical trials and provide greater insight into the various dimensions of the clinical effect of interventions included in the trials. Various strategies have been proposed for using longitudinal findings to elucidate intervention effects on clinical outcomes such as survival. The application of specifically Bayesian approaches for constructing models that address longitudinal and survival outcomes explicitly has been recently addressed in the literature. We review currently available methods for carrying out joint analyses, including issues of implementation and interpretation, identify software tools that can be used to carry out the necessary calculations, and review applications of the methodology.

Keywords: time-dependent; random effects; software; applications

References I

Crowther, M. J., Abrams, K. R., and Lambert, P. C. Joint modeling of longitudinal and survival data. *Stata J*, 13 (1):165–184, 2013a.

Crowther, M. J., Lambert, P. C., and Abrams, K. R. Adjusting for measurement error in baseline prognostic biomarkers included in a time-to-event analysis: A joint modelling approach. *BMC Med Res Methodol*, 13 (146), 2013b.

Dantan, E., Joly, P., Dartigues, J.-F., and Jacqmin-Gadda, H. Joint model with latent state for longitudinal and multistate data. *Biostatistics*, 12(4):723–736, Oct 2011.

Faucett, C. L. and Thomas, D. C. Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Stat Med*, 15(15):1663–1685, 1996.

Gould, A. L., Boye, M. E., Crowther, M. J., Ibrahim, J. G., Quartey, G., Micallef, S., and Bois, F. Y. Joint modeling of survival and longitudinal non-survival data: current methods and issues. Report of the DIA Bayesian joint modeling working group. *Stat Med*, 2014.

Henderson, R., Diggle, P., and Dobson, A. Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1(4):465–480, 2000.

Jutte, D. P., Roos, L. L., and Brownell, M. D. Administrative record linkage as a tool for public health research. *Annu Rev Public Health*, 32:91–108, 2011.

Li, J., Humphreys, K., Eriksson, L., Edgren, G., Czene, K., and Hall, P. Mammographic density reduction is a prognostic marker of response to adjuvant tamoxifen therapy in postmenopausal patients with breast cancer. *J Clin Oncol*, 31(18):2249–2256, Jun 2013.

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

Pinheiro, J. C. and Bates, D. M. Approximations to the log-likelihood function in the nonlinear mixed-effects model. *J Comput Graph Statist*, 4(1):pp. 12–35, 1995.

References II

Proust-Lima, C. and Taylor, J. M. G. 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.

Proust-Lima, C., Joly, P., Dartigues, J.-F., and Jacqmin-Gadda, H. Joint modelling of multivariate longitudinal outcomes and a time-to-event: A nonlinear latent class approach. *Computational Statistics & Data Analysis*, 53(4):1142 – 1154, 2009.

Proust-Lima, C., Sene, M., Taylor, J. M., and Jacqmin-Gadda, H. Joint latent class models for longitudinal and time-to-event data: A review. *Stat Methods Med Res*, Apr 2012.

Thiébaut, A. C. M. and Bénichou, J. Choice of time-scale in Cox's model analysis of epidemiologic cohort data: a simulation study. *Stat Med*, 23(24):3803–3820, 2004.

van den Hout, A. and Muniz-Terrera, G. Joint models for discrete longitudinal outcomes in ageing research. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 2015.

Wolbers, M., Babiker, A., Sabin, C., Young, J., Dorrucci, M., Chêne, G., Mussini, C., Porter, K., Bucher, H. C., and CASCADE. Pretreatment CD4 cell slope and progression to AIDS or death in HIV-infected patients initiating antiretroviral therapy–the CASCADE collaboration: a collaboration of 23 cohort studies. *PLoS Med*, 7(2):e1000239, 2010.

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