Bayesian joint models for multiple longitudinal biomarkers and event-time data: methods and software development

Sam Brilleman[1,2], Michael J. Crowther[3], Margarita Moreno-Betancur[1,2,4], Rory Wolfe[1,2]

Australian Statistical Conference

Canberra, Australia

5-9th December 2016

[1] Monash University, Australia

[2] Victorian Centre for Biostatistics (ViCBiostat)

[3] University of Leicester, UK

[4] Murdoch Childrens Research Institute, Australia

Background

What is joint modelling?

- The joint estimation of distinct regression models which, traditionally, we would have estimated separately
  - One or more longitudinal (mixed effects) models
    - each for a repeatedly measured clinical marker, e.g. systolic blood pressure
  - A survival or time-to-event (proportional hazards) model
    - for the time to an event, e.g. time-to-death, time-to-stroke

Background

Why use joint modelling?

- We want to know whether the longitudinal marker is associated with the risk of the event
  - e.g. how is time-varying SBP associated with the risk of death?
  - can actually consider association between the event risk and **any aspect** of the longitudinal trajectory (e.g. slope)
  - can allow for **measurement error** in the marker
  - can allow for **discrete-time** measurement of the marker
- And possibly other reasons…
  - e.g. dynamic predictions, separating out “direct” and “indirect” effects of treatment, adjusting for informative dropout

$y_{ik}(t)$ is the value at time t of the
k^{th} longitudinal marker ($k = 1, \dots, K$)
for the $i^{\text{ th}}$ individual ($i = 1, \dots, N$)
T_i is “true” event time, C_i is the censoring time
$T_i^* = \min(T_i, C_i)$ and $d_i = I(T_i \le C_i)$

Joint model specification

Longitudinal submodel

$y_{ik}(t)$ follows a distribution in the exponential family with expected value $\mu_{ik}(t)$ and

$$\eta_{ik}(t) = g_k\big(\mu_{ik}(t)\big) = \boldsymbol{x}'_{\boldsymbol{ik}}(t)\boldsymbol{\beta}_{\boldsymbol{k}} + \boldsymbol{z}'_{\boldsymbol{ik}}(t)\boldsymbol{b}_{\boldsymbol{ik}}$$

$$\begin{bmatrix} \boldsymbol{b}_{\boldsymbol{i1}} \\ \vdots \\ \boldsymbol{b}_{\boldsymbol{iK}} \end{bmatrix} = \boldsymbol{b}_{\boldsymbol{i}} \sim N(0, \boldsymbol{\Sigma})$$

Event submodel

$$h_i(t) = h_0(t) \exp\left(\boldsymbol{w}'_{\boldsymbol{i}}(t)\boldsymbol{\gamma} + \sum_{k=1}^{K} \sum_{q=1}^{Q_k} \alpha_{kq} f_{kq}(\eta_{ik}(t), \mu_{ik}(t), \boldsymbol{\beta}_k, \boldsymbol{b}_{\boldsymbol{ik}}) \right)$$

Association structures

$y_{ik}(t)$ is the value at time t of the
k^{th} longitudinal marker ($k = 1, \dots, K$)
for the i^{th} individual ($i = 1, \dots, N$)
T_i is “true” event time, C_i is the censoring time
$T_i^* = \min(T_i, C_i)$ and $d_i = I(T_i \le C_i)$

Longitudinal submodel

$y_{ik}(t)$ follows a distribution in the exponential family with expected value $\mu_{ik}(t)$ and

$$\eta_{ik}(t) = g_k\big(\mu_{ik}(t)\big) = \boldsymbol{x}_{\boldsymbol{ik}}'(t)\boldsymbol{\beta}_{\boldsymbol{k}} + \boldsymbol{z}_{\boldsymbol{ik}}'(t)\boldsymbol{b}_{\boldsymbol{ik}}$$

$$\begin{bmatrix} \boldsymbol{b}_{\boldsymbol{i1}} \\ \vdots \\ \boldsymbol{b}_{\boldsymbol{iK}} \end{bmatrix} = \boldsymbol{b}_{\boldsymbol{i}} \sim N(0, \boldsymbol{\Sigma})$$

Event submodel

$$h_i(t) = h_0(t)\exp\left(\boldsymbol{w}_{\boldsymbol{i}}'(t)\boldsymbol{\gamma} + \sum_{k=1}^{K}\sum_{q=1}^{Q_k}\alpha_{kq} f_{kq}(\eta_{ik}(t), \mu_{ik}(t), \boldsymbol{\beta}_k, \boldsymbol{b}_{\boldsymbol{ik}})\right)$$

Association structures

$$\boxed{f_{kq}(\eta_{ik}(t), \mu_{ik}(t), \boldsymbol{\beta_k}, \boldsymbol{b_{ik}}) = ?}$$

$\eta_{ik}(t)$ → Value of the linear predictor at time t

$\mu_{ik}(t)$ → Expected value of the marker at time t

$\frac{d\mu_{ik}(t)}{dt}$ → Rate of change in the marker (i.e. slope) at time t

$\int_0^t \mu_{ik}(s)\, ds$ → Area under the marker trajectory (e.g. cumulative dose) up to time t

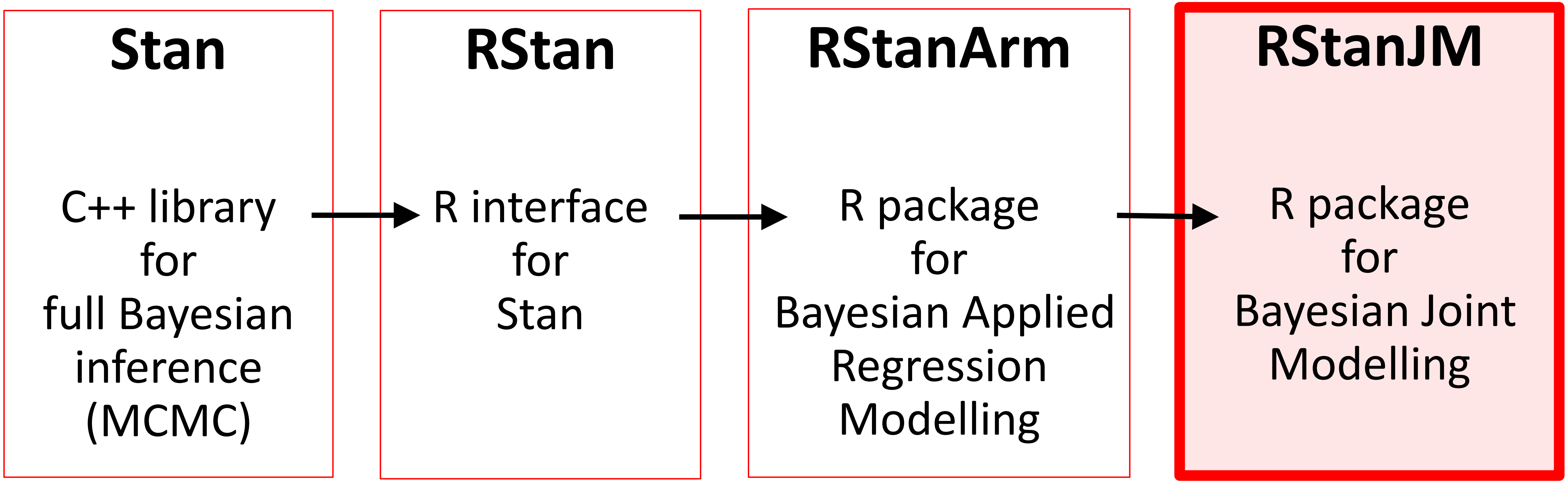
Joint model likelihood

Likelihood function:

$$p(\boldsymbol{y}_{\boldsymbol{i1}},\dots,\boldsymbol{y}_{\boldsymbol{iK}},T_i,d_i|\boldsymbol{b}_{\boldsymbol{i}},\boldsymbol{\theta}) = \int_{-\infty}^{\infty}\left(\prod_{k=1}^{K}\underbrace{\prod_{j=1}^{n_{ik}} p\big(y_{ik}(t_{ijk})\big|\boldsymbol{b}_{\boldsymbol{i}},\boldsymbol{\theta}_{\boldsymbol{y}_{\boldsymbol{k}}}\big)}_{k\text{th longitudinal submodel}}\right)\underbrace{p(T_i,d_i|\boldsymbol{b}_{\boldsymbol{i}},\boldsymbol{\theta}_{\boldsymbol{T}})}_{\text{event submodel}}\,\underbrace{p(\boldsymbol{b}_{\boldsymbol{i}}|\boldsymbol{\theta}_{\boldsymbol{b}})}_{\text{random effects model}}\,\mathrm{d}\boldsymbol{b}_{\boldsymbol{i}}$$

- Assumes **conditional independence**, that is, conditional on $\boldsymbol{b}_{\boldsymbol{i}}$ the distinct longitudinal and event processes are independent
  - requires we specify the model correctly, including the **“association structure”**
- Time-dependence in the event likelihood poses an additional computational burden

Bayesian joint models via Stan



Stan

C++ library
for
full Bayesian
inference
(MCMC)

RStan

R interface
for
Stan

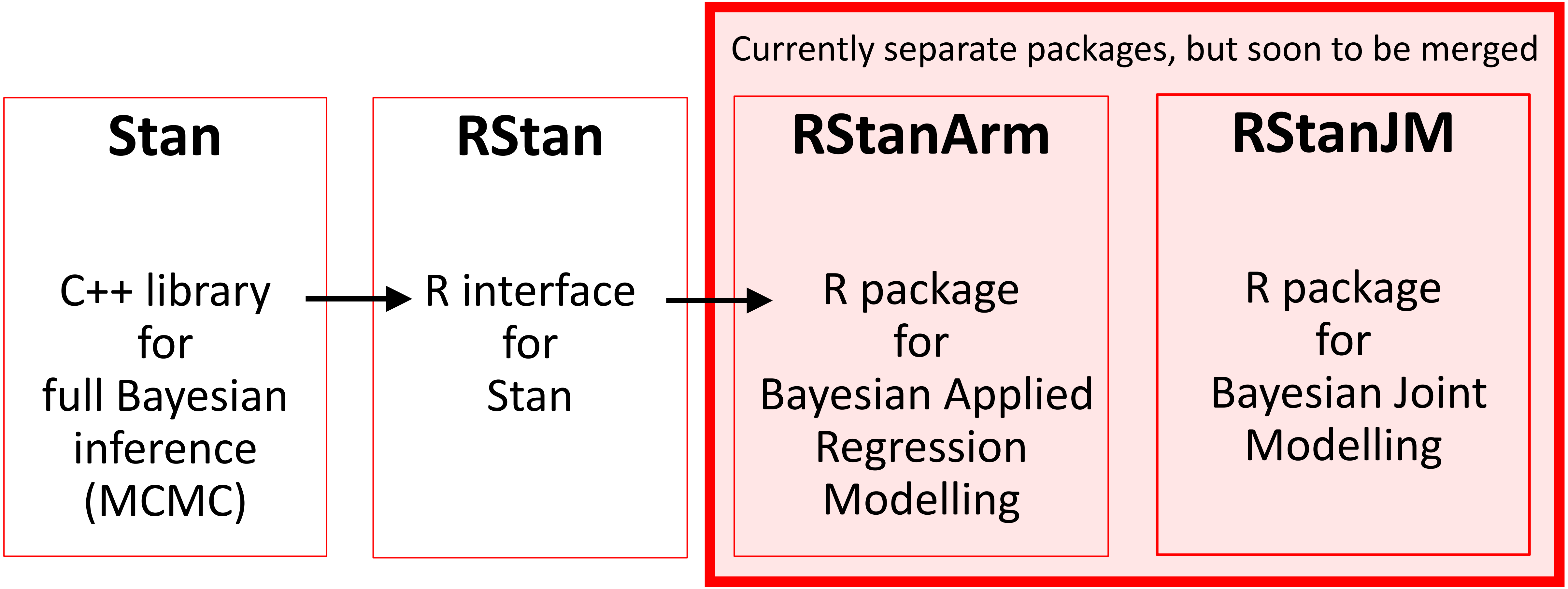
RStanArm

R package
for
Bayesian Applied
Regression
Modelling

RStanJM

R package
for
Bayesian Joint
Modelling

Bayesian joint models via Stan



Currently separate packages, but soon to be merged

Stan

C++ library for full Bayesian inference (MCMC)

RStan

R interface for Stan

RStanArm

R package for Bayesian Applied Regression Modelling

RStanJM

R package for Bayesian Joint Modelling

Bayesian joint models via Stan

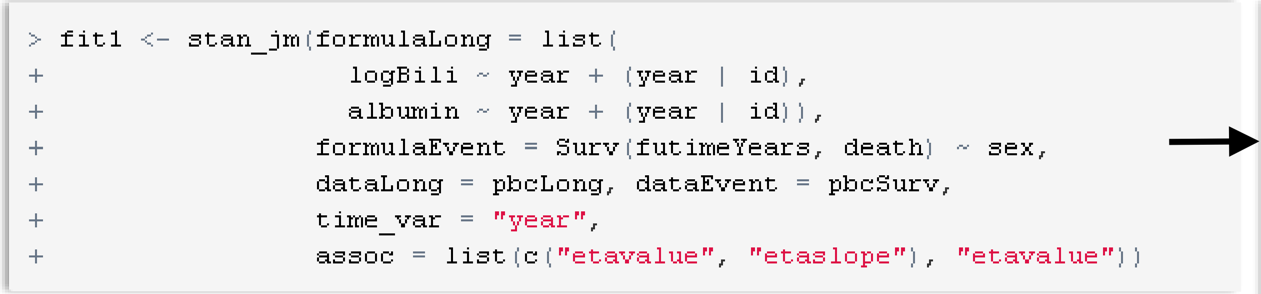
- Development version currently available as a stand-alone package ‘rstanjm’
  - https://github.com/sambrilleman/rstanjm
- Association structures
  - current value or slope (of linear predictor or mean)
  - shared random effects (optionally including fixed effect component)
- Variety of prior distributions
  - Regression coefficients: normal, student t, Cauchy, and horseshoe (shrinkage) priors
  - Novel decomposition of covariance matrix for the random effects
- Variety of link functions and error distributions
  - Incl. normal, binomial, Poisson, negative binomial, gamma
- Baseline hazard
  - Weibull, piecewise constant, or B-splines approximation

Example

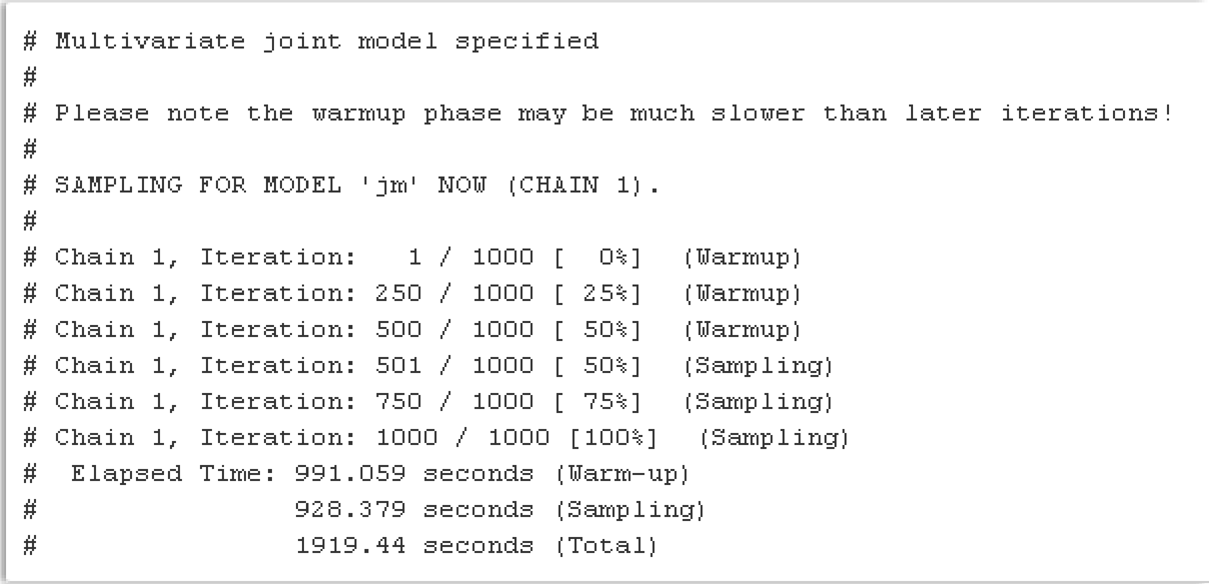
- Data: Mayo Clinic’s primary biliary cirrhosis (“PBC”) data

- Longitudinal submodels:
    - Outcomes: log serum bilirubin, albumin
    - Linear mixed model w/ random intercept and random linear slope

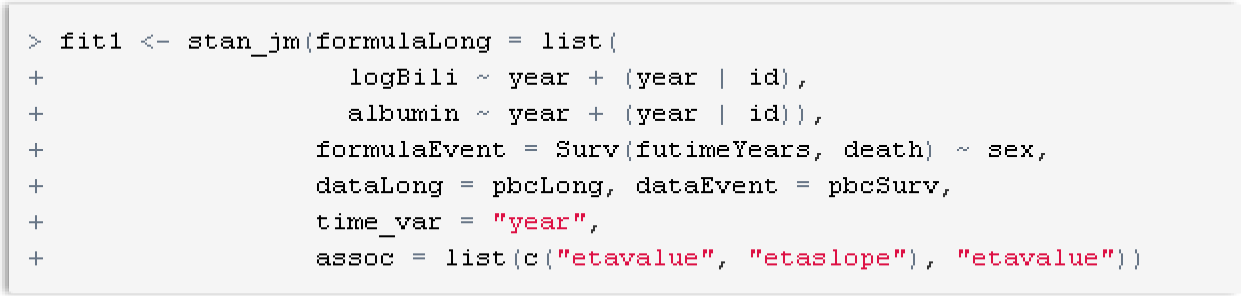
- Event submodel
    - Time-fixed covariate: gender
    - Association structure: current value and slope (bilirubin), current value (albumin)
    - Weibull baseline hazard



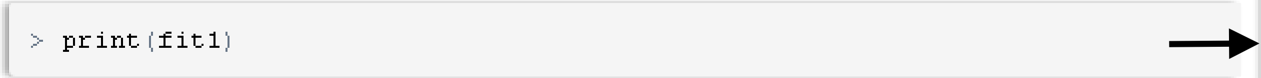
```
> fit1 <- stan_jm(formulaLong = list(
+                   logBili ~ year + (year | id),
+                   albumin ~ year + (year | id)),
+                 formulaEvent = Surv(futimeYears, death) ~ sex,
+                 dataLong = pbcLong, dataEvent = pbcSurv,
+                 time_var = "year",
+                 assoc = list(c("etavalue", "etaslope"), "etavalue"))
```



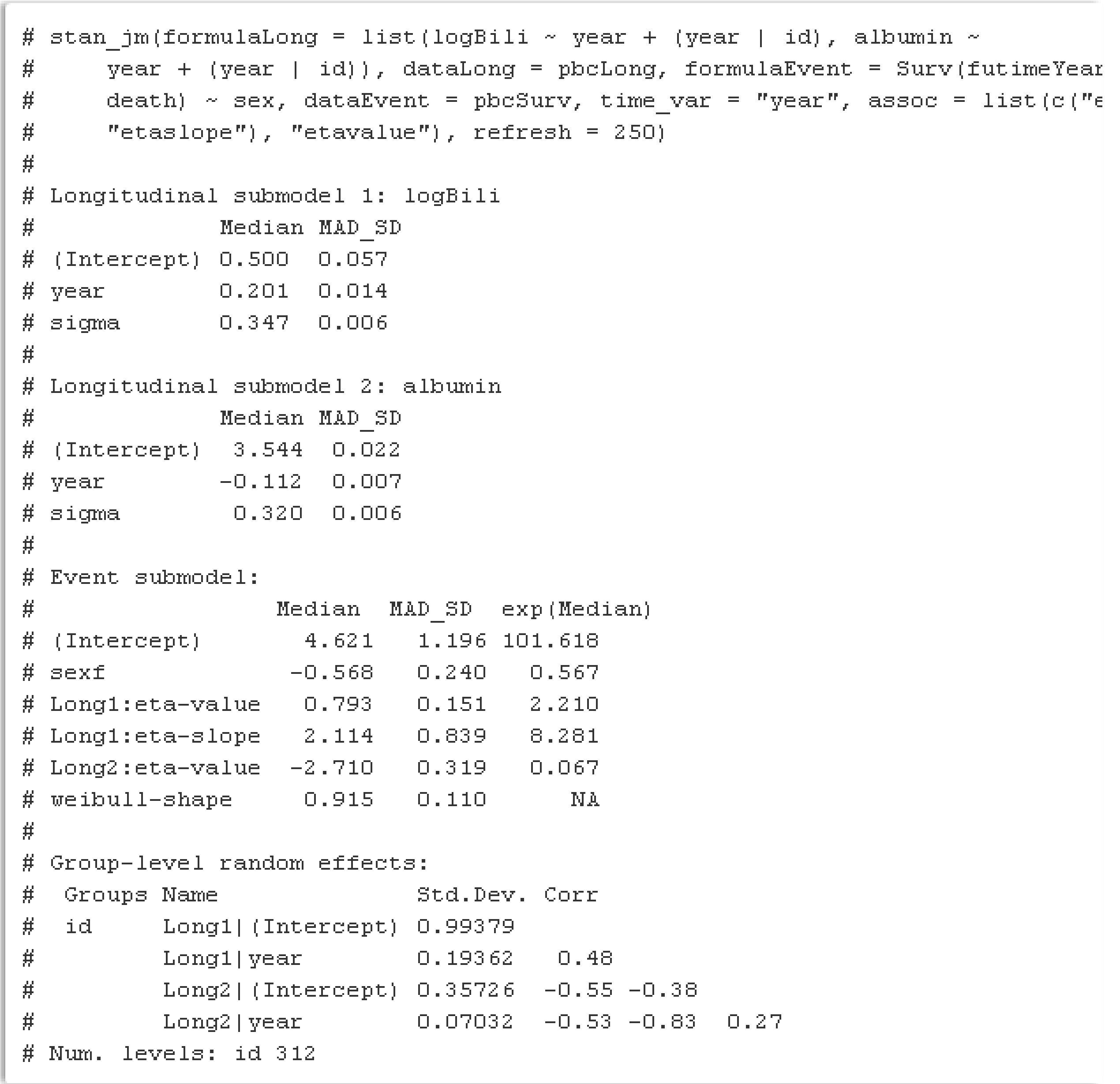
```
# Multivariate joint model specified
#
# Please note the warmup phase may be much slower than later iterations!
#
# SAMPLING FOR MODEL 'jm' NOW (CHAIN 1).
#
# Chain 1, Iteration:    1 / 1000 [  0%]  (Warmup)
# Chain 1, Iteration: 250 / 1000 [ 25%]  (Warmup)
# Chain 1, Iteration: 500 / 1000 [ 50%]  (Warmup)
# Chain 1, Iteration: 501 / 1000 [ 50%]  (Sampling)
# Chain 1, Iteration: 750 / 1000 [ 75%]  (Sampling)
# Chain 1, Iteration: 1000 / 1000 [100%]  (Sampling)
#  Elapsed Time: 991.059 seconds (Warm-up)
#                928.379 seconds (Sampling)
#                1919.44 seconds (Total)
```



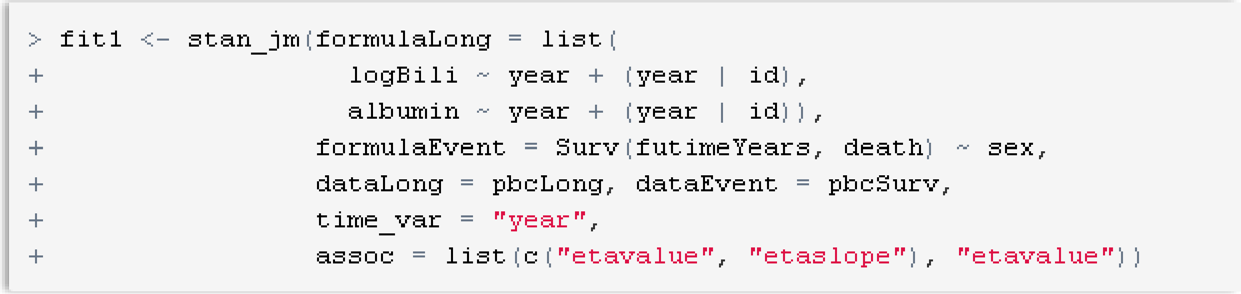
```
> fit1 <- stan_jm(formulaLong = list(
+                   logBili ~ year + (year | id),
+                   albumin ~ year + (year | id)),
+                 formulaEvent = Surv(futimeYears, death) ~ sex,
+                 dataLong = pbcLong, dataEvent = pbcSurv,
+                 time_var = "year",
+                 assoc = list(c("etavalue", "etaslope"), "etavalue"))
```



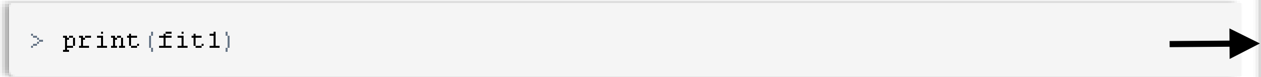
```
> print(fit1)
```



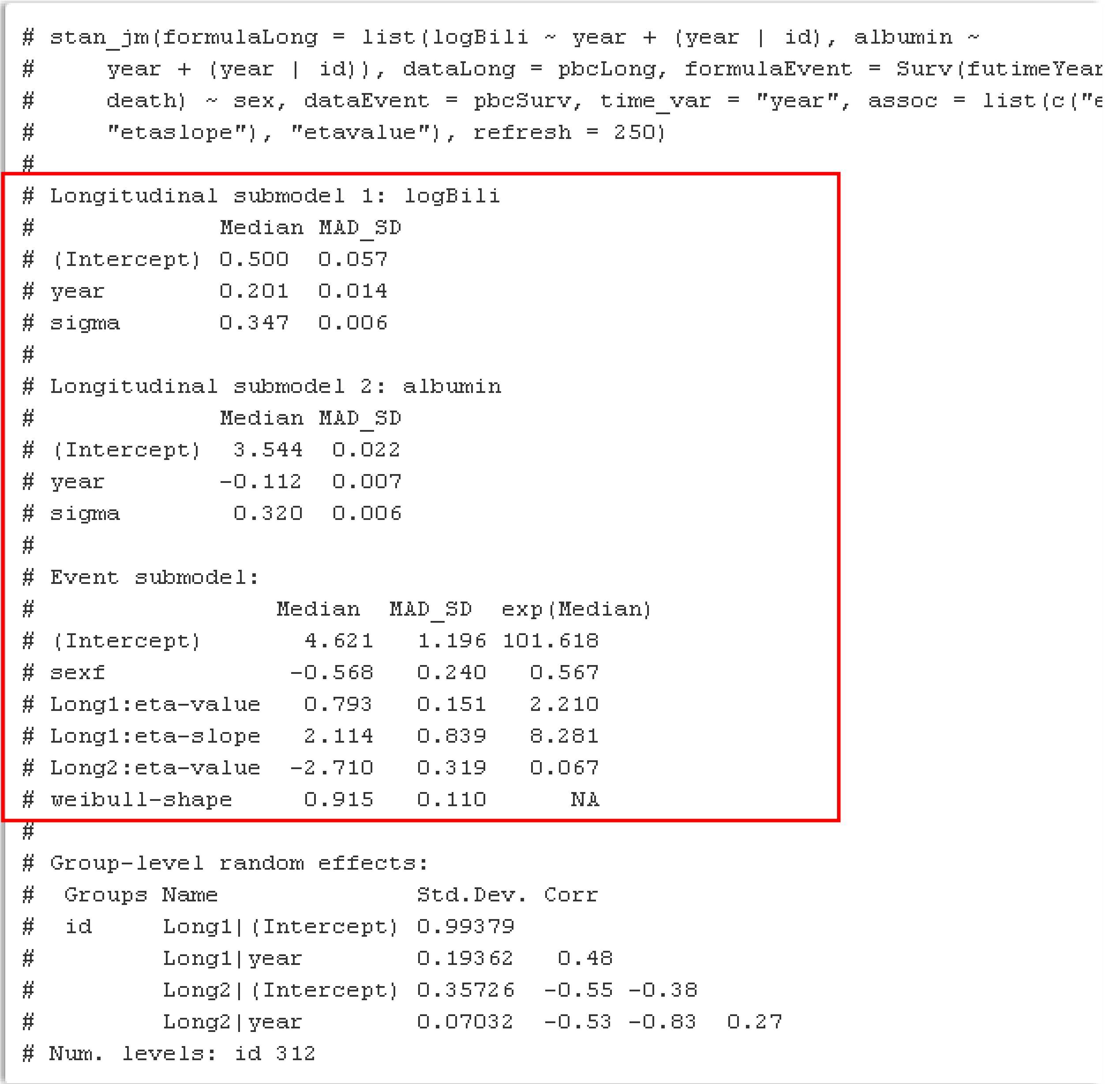
```
# stan_jm(formulaLong = list(logBili ~ year + (year | id), albumin ~
#     year + (year | id)), dataLong = pbcLong, formulaEvent = Surv(futimeYear
#     death) ~ sex, dataEvent = pbcSurv, time_var = "year", assoc = list(c("e
#     "etaslope"), "etavalue"), refresh = 250)
#
# Longitudinal submodel 1: logBili
#             Median MAD_SD
# (Intercept) 0.500  0.057
# year        0.201  0.014
# sigma       0.347  0.006
#
# Longitudinal submodel 2: albumin
#             Median MAD_SD
# (Intercept)  3.544  0.022
# year        -0.112  0.007
# sigma        0.320  0.006
#
# Event submodel:
#                 Median  MAD_SD  exp(Median)
# (Intercept)       4.621   1.196 101.618
# sexf             -0.568   0.240   0.567
# Long1:eta-value   0.793   0.151   2.210
# Long1:eta-slope   2.114   0.839   8.281
# Long2:eta-value  -2.710   0.319   0.067
# weibull-shape     0.915   0.110      NA
#
# Group-level random effects:
#  Groups Name              Std.Dev. Corr
#  id     Long1|(Intercept) 0.99379
#         Long1|year        0.19362   0.48
#         Long2|(Intercept) 0.35726  -0.55 -0.38
#         Long2|year        0.07032  -0.53 -0.83  0.27
# Num. levels: id 312
```



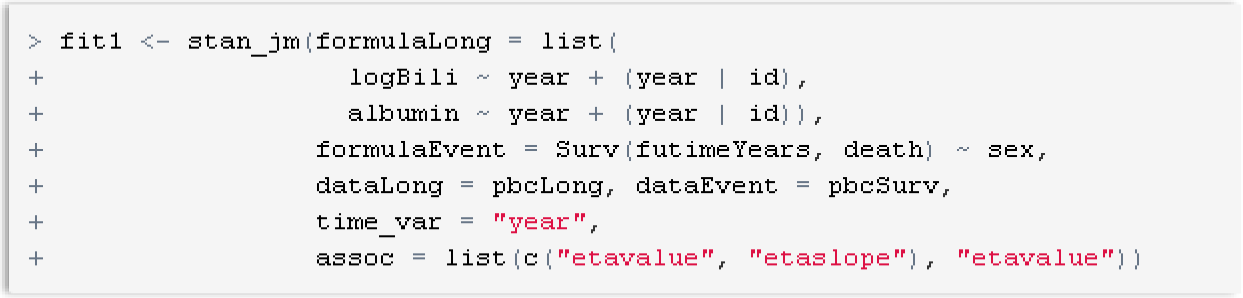
```
> fit1 <- stan_jm(formulaLong = list(
+                    logBili ~ year + (year | id),
+                    albumin ~ year + (year | id)),
+                  formulaEvent = Surv(futimeYears, death) ~ sex,
+                  dataLong = pbcLong, dataEvent = pbcSurv,
+                  time_var = "year",
+                  assoc = list(c("etavalue", "etaslope"), "etavalue"))
```



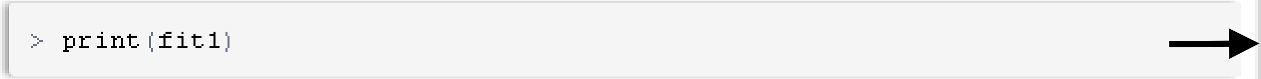
```
> print(fit1)
```



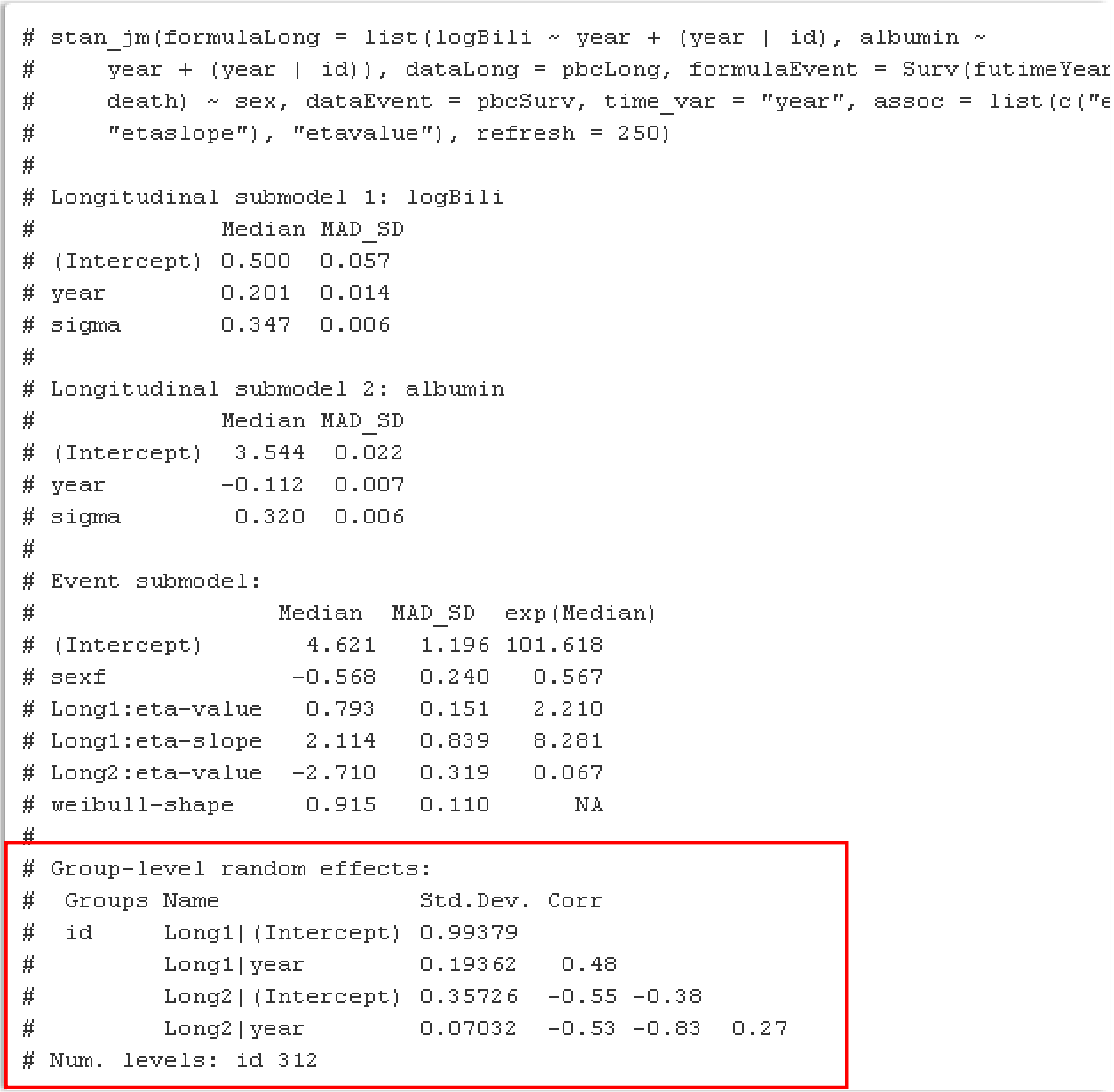
```
# stan_jm(formulaLong = list(logBili ~ year + (year | id), albumin ~
#     year + (year | id)), dataLong = pbcLong, formulaEvent = Surv(futimeYear
#     death) ~ sex, dataEvent = pbcSurv, time_var = "year", assoc = list(c("e
#     "etaslope"), "etavalue"), refresh = 250)
#
# Longitudinal submodel 1: logBili
#             Median MAD_SD
# (Intercept) 0.500  0.057
# year        0.201  0.014
# sigma       0.347  0.006
#
# Longitudinal submodel 2: albumin
#             Median MAD_SD
# (Intercept)  3.544  0.022
# year        -0.112  0.007
# sigma        0.320  0.006
#
# Event submodel:
#                 Median  MAD_SD  exp(Median)
# (Intercept)       4.621   1.196 101.618
# sexf             -0.568   0.240   0.567
# Long1:eta-value   0.793   0.151   2.210
# Long1:eta-slope   2.114   0.839   8.281
# Long2:eta-value  -2.710   0.319   0.067
# weibull-shape     0.915   0.110      NA
#
# Group-level random effects:
#  Groups Name              Std.Dev. Corr
#  id     Long1|(Intercept) 0.99379
#         Long1|year        0.19362   0.48
#         Long2|(Intercept) 0.35726  -0.55 -0.38
#         Long2|year        0.07032  -0.53 -0.83  0.27
# Num. levels: id 312
```



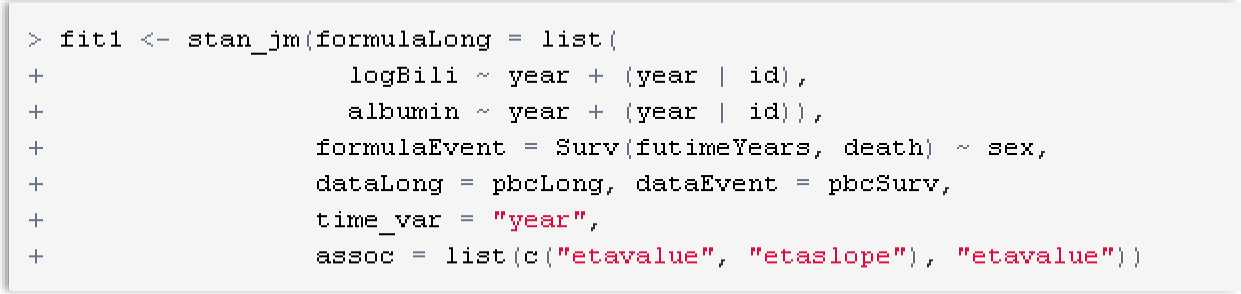
```
> fit1 <- stan_jm(formulaLong = list(
+                   logBili ~ year + (year | id),
+                   albumin ~ year + (year | id)),
+                 formulaEvent = Surv(futimeYears, death) ~ sex,
+                 dataLong = pbcLong, dataEvent = pbcSurv,
+                 time_var = "year",
+                 assoc = list(c("etavalue", "etaslope"), "etavalue"))
```



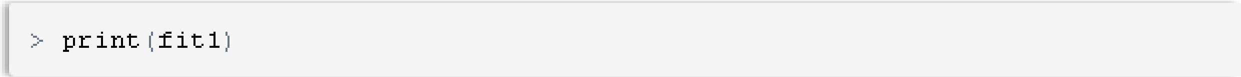
```
> print(fit1)
```



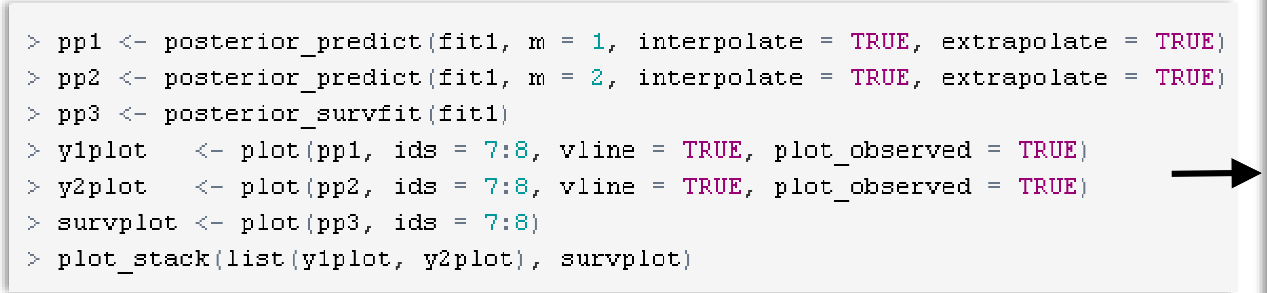
```
# stan_jm(formulaLong = list(logBili ~ year + (year | id), albumin ~
#     year + (year | id)), dataLong = pbcLong, formulaEvent = Surv(futimeYear
#     death) ~ sex, dataEvent = pbcSurv, time_var = "year", assoc = list(c("e
#     "etaslope"), "etavalue"), refresh = 250)
#
# Longitudinal submodel 1: logBili
#              Median MAD_SD
# (Intercept) 0.500  0.057
# year        0.201  0.014
# sigma       0.347  0.006
#
# Longitudinal submodel 2: albumin
#              Median MAD_SD
# (Intercept)  3.544  0.022
# year        -0.112  0.007
# sigma        0.320  0.006
#
# Event submodel:
#                 Median  MAD_SD  exp(Median)
# (Intercept)       4.621   1.196 101.618
# sexf             -0.568   0.240   0.567
# Long1:eta-value   0.793   0.151   2.210
# Long1:eta-slope   2.114   0.839   8.281
# Long2:eta-value  -2.710   0.319   0.067
# weibull-shape     0.915   0.110      NA
#
# Group-level random effects:
#  Groups Name              Std.Dev. Corr
#  id     Long1|(Intercept) 0.99379
#         Long1|year        0.19362   0.48
#         Long2|(Intercept) 0.35726  -0.55 -0.38
#         Long2|year        0.07032  -0.53 -0.83  0.27
# Num. levels: id 312
```



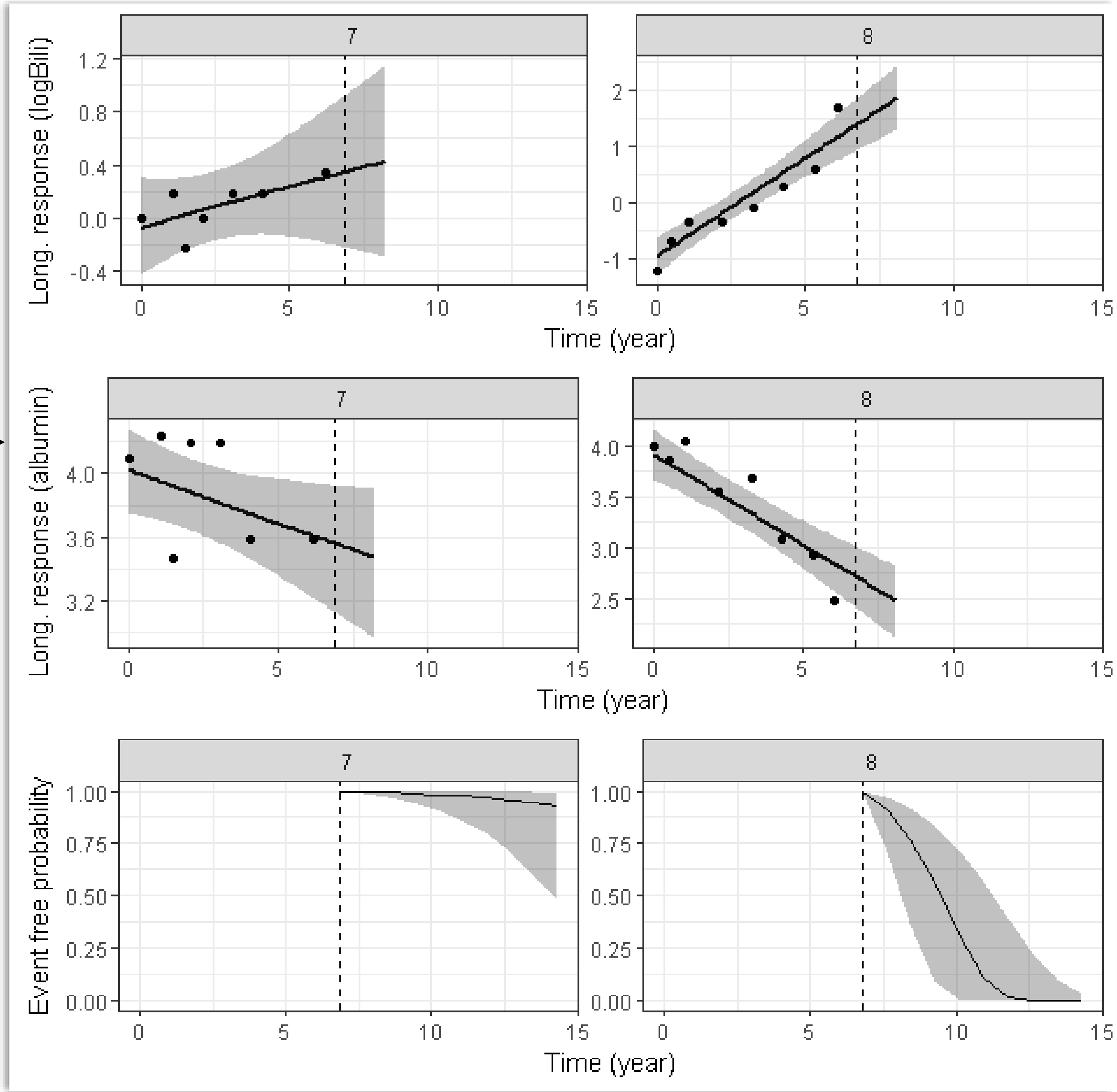
```
> fit1 <- stan_jm(formulaLong = list(
+                     logBili ~ year + (year | id),
+                     albumin ~ year + (year | id)),
+                   formulaEvent = Surv(futimeYears, death) ~ sex,
+                   dataLong = pbcLong, dataEvent = pbcSurv,
+                   time_var = "year",
+                   assoc = list(c("etavalue", "etaslope"), "etavalue"))
```

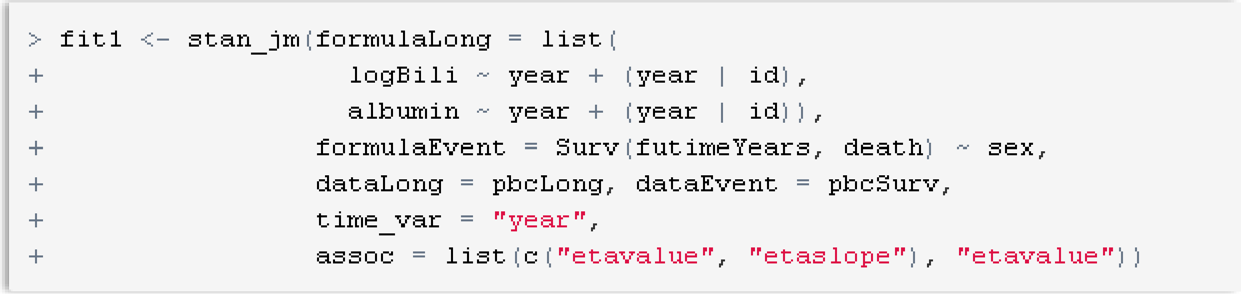


```
> print(fit1)
```

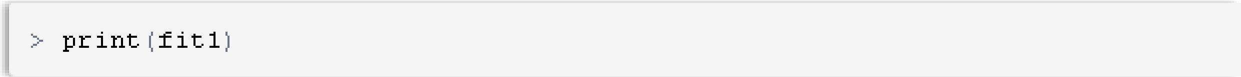


```
> pp1 <- posterior_predict(fit1, m = 1, interpolate = TRUE, extrapolate = TRUE)
> pp2 <- posterior_predict(fit1, m = 2, interpolate = TRUE, extrapolate = TRUE)
> pp3 <- posterior_survfit(fit1)
> y1plot   <- plot(pp1, ids = 7:8, vline = TRUE, plot_observed = TRUE)
> y2plot   <- plot(pp2, ids = 7:8, vline = TRUE, plot_observed = TRUE)
> survplot <- plot(pp3, ids = 7:8)
> plot_stack(list(y1plot, y2plot), survplot)
```

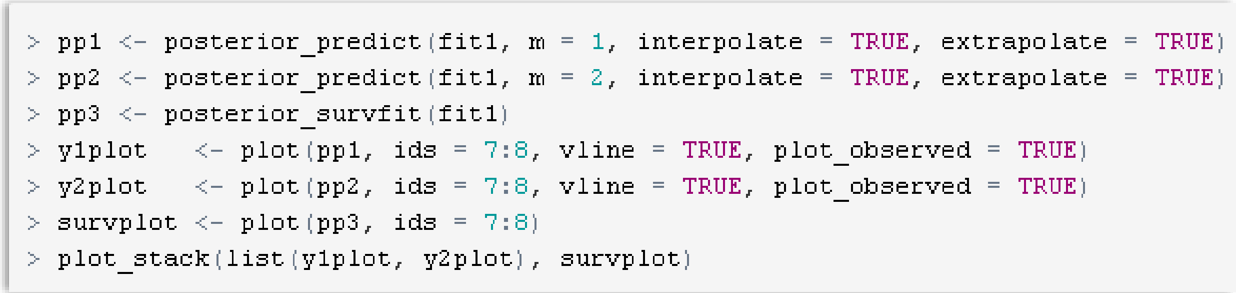




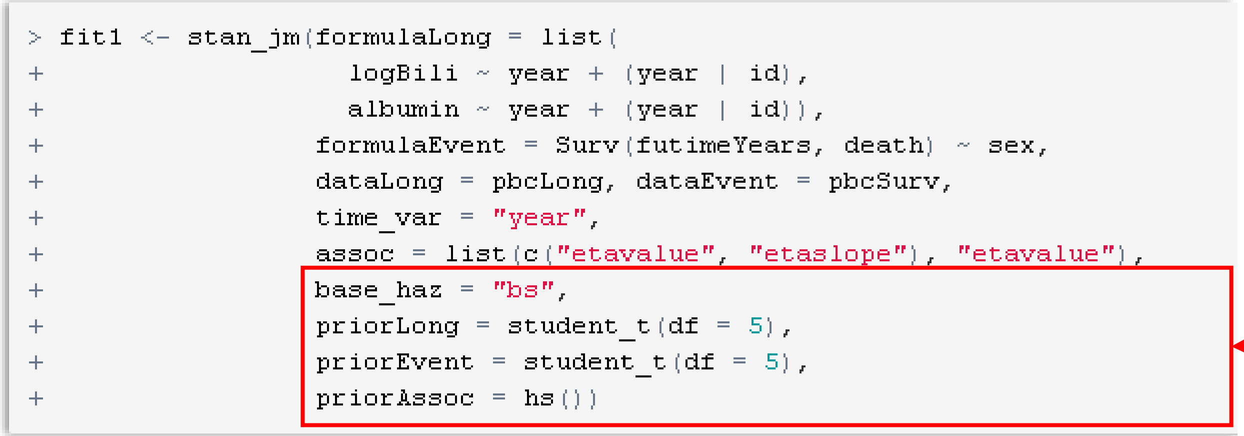
```
> fit1 <- stan_jm(formulaLong = list(
+                   logBili ~ year + (year | id),
+                   albumin ~ year + (year | id)),
+                 formulaEvent = Surv(futimeYears, death) ~ sex,
+                 dataLong = pbcLong, dataEvent = pbcSurv,
+                 time_var = "year",
+                 assoc = list(c("etavalue", "etaslope"), "etavalue"))
```



```
> print(fit1)
```



```
> pp1 <- posterior_predict(fit1, m = 1, interpolate = TRUE, extrapolate = TRUE)
> pp2 <- posterior_predict(fit1, m = 2, interpolate = TRUE, extrapolate = TRUE)
> pp3 <- posterior_survfit(fit1)
> y1plot   <- plot(pp1, ids = 7:8, vline = TRUE, plot_observed = TRUE)
> y2plot   <- plot(pp2, ids = 7:8, vline = TRUE, plot_observed = TRUE)
> survplot <- plot(pp3, ids = 7:8)
> plot_stack(list(y1plot, y2plot), survplot)
```



```
> fit1 <- stan_jm(formulaLong = list(
+                   logBili ~ year + (year | id),
+                   albumin ~ year + (year | id)),
+                 formulaEvent = Surv(futimeYears, death) ~ sex,
+                 dataLong = pbcLong, dataEvent = pbcSurv,
+                 time_var = "year",
+                 assoc = list(c("etavalue", "etaslope"), "etavalue"),
+                 base_haz = "bs",
+                 priorLong = student_t(df = 5),
+                 priorEvent = student_t(df = 5),
+                 priorAssoc = hs())
```

Can easily change priors or baseline hazard

Thank you

- My PhD supervisors: Rory Wolfe, Margarita Moreno-Betancur, Michael Crowther, John Carlin
- My PhD funders: NHMRC and Victorian Centre for Biostatistics (ViCBiostat)
- Staff from ViCBiostat ☺
- Ben Goodrich and Jonah Gabry (authors of RStanArm)

MONASH University

ViCBiostat