Joint Modelling of Longitudinal Measurements and Survival Outcomes

Peter Diggle

Medical Statistics Unit, Lancaster University
and
Department of Biostatistics, Johns Hopkins University

Copenhagen, April 2006

Outline

- Joint modelling:
  - what is it?
  - why do it?
- Modelling longitudinal measurements
- Modelling survival outcomes
- Joint modelling
  - transformation models
  - random effects models
- Examples
- Conclusions

Joint modelling: what is it?

- Subjects $i = 1, ..., m$.
- Longitudinal measurements Y_{ij} at times t_{ij}.
- Times-to-event S_i (possibly censored).
- Baseline covariates x_i.
- Parameters θ.

$$[Y, S|x, \theta]$$

AIDS data

- Data from RCT of three different drug regimes for HIV+ patients.

  $1 =$ zidovudine, $600mg$

  $2 =$didanosine, $500mg$

  $3 =$ didanosine, $750mg$

  Total $n = 913$ subjects.

- $S =$ time of progression to AIDS or death

  $Y =$ time-sequence of CD4 measurements (weeks 0, 2, 6, 12, 18, 24)

- $S > 24$ for most subjects (long-term follow-up)

- 334 observed event-times, 579 censored

Schizophrenia data

- Data from placebo-controlled RCT of drug treatments for schizophrenia:
  - Placebo
  - Haloperidol (standard)
  - Risperidone (novel)

  Total $n = 517$ subjects.
- Y = PANSS measurement (weeks -1, 0, 1, 2, 4, 6, 8)
- S = dropout time
- High dropout rates:

week	-1	0	1	2	4	6	8
missing	0	3	9	70	122	205	251
proportion	0.00	0.01	0.02	0.14	0.24	0.40	0.49

- Dropout rate also treatment-dependent ($P > H > R$)

Heart surgery data

- RCT to compare two types of artificial heart-valve
  - homograft
  - stentless
- Y = time-sequence of left-ventricular-mass-index (LVMI)
- S = time of death

Is a patient's longitudinal LVMI profile predictive of their survival prognosis?

Joint modelling: why do it?

To analyse survival time S, whilst exploiting correlation with an imperfectly measured, time-varying risk-factor Y

Example: AIDS data

- interest is in time to progression/death
- but long latency period implies heavy censoring
- hence, joint modelling improves inferences about marginal distribution $[S]$

To analyse a longitudinal outcome measure Y with potentially informative dropout at time S

Example: Schizophrenia data

- **interest is reducing mean PANSS score**
- **but informative dropout process would imply that modelling only $[Y]$ may be misleading**

Because the relationship between Y and S is of intrinsic interest

Example: heart surgery data

- **long-term build-up of left-ventricular muscle mass may increase hazard for fatal heart-attack**
- **hence, interested in modelling relationship between survival and subject-level LVMI**

Scientific goal affects choice of statistical model/method?

Modelling longitudinal measurements

The Gaussian linear model

- $(Y_{ij}, t_{ij}) : j = 1, ..., n_i;\ i = 1, ..., m \qquad \mu_{ij} = \mathrm{E}[Y_{ij}]$
- $Y_i = (Y_{i1}, ..., Y_{in_i}) \qquad Y = (Y_1, ..., Y_m)$
- $t_i = (t_{i1}, ..., t_{in_i}) \qquad t = (t_1, ..., t_m)$

$$Y \sim \mathrm{MVN}\{X\beta, V(\phi)\})$$

$$V(\phi) = \begin{bmatrix} V_1 & 0 & \dots & 0 \\ 0 & V_2 & \ddots & \vdots \\ \vdots & \ddots & & 0 \\ 0 & \dots & 0 & V_m \end{bmatrix}$$

Modelling longitudinal measurements

Specifying the covariance structure

$$Y_{ij} = \mu_{ij} + U_i + W_i(t_{ij}) + Z_{ij}$$

Corresponds to within-subject variance matrices

$$V_i = \nu^2 J + \sigma^2 R(t_i) + \tau^2 I$$

where $R(t_i)$ has elements $R_{jk} = \rho(|t_{ij} - t_{ik}|)$.

Modelling longitudinal measurements

The variogram

$$V(u) = \frac{1}{2}\mathrm{Var}\{Y(t) - Y(t-u)\}$$

- useful data-analytic tool for irregularly spaced, incomplete data
- theoretical form for model on previous slide

$$V(u) = \tau^2 + \sigma^2\{1 - \rho(u)\}$$

Modelling survival outcomes

Fundamental tool is the hazard function,

$$h(s) = f(s)/\{1 - F(s)\}$$

Modelling strategies

- **Proportional hazards**

  $h_i(s) = h_0(s)\theta_i \quad \theta_i = \exp(\mathrm{x}_i'\beta)$

- **Accelerated life**

  $F_i(s) = F_0(\theta_i s) \quad \theta_i = \exp(\mathrm{x}_i'\beta)$

- **Frailty**

  $h_i(s) = h_0(s)\theta_i U_i \quad \theta_i = \exp(\mathrm{x}_i'\beta) \quad U_i \sim G(u)$

Modelling strategies (continued)

- **Parametric**

  $S \sim f(s;\theta) \quad \theta_i = \exp(\mathrm{x}_i'\beta)$

  - $S \sim \mathrm{Gamma}(\lambda, \kappa)$: proportional hazards (κ known)
  - $\log S \sim \mathrm{N}(\mu, \sigma^2)$: accelerated life
  - $S \sim \mathrm{Weibull}(\lambda, \delta)$: proportional hazards and accelerated life

Joint modelling

Considerations to inform choice of approach

- focus on questions of primary scientific interest
- interpretability of model parameters
- statistical efficiency
- diagnostic checks for assumptions about effects of primary interest
- robustness to departures from assumptions about effects not of primary interest
- ease of implementation
- reduction to standard methods when there is no association

Joint modelling

Transformation models

$$(\log S, Y) \sim \text{MVN}(\mu, \Sigma)$$

- $\mu = (\mu_S, \mu_Y)$
- $\Sigma = \begin{bmatrix} \sigma^2 & \gamma' \\ \gamma & V(\phi) \end{bmatrix}$
- subjects provide independent replicates of $(\log S, Y)$

Transformation models: the likelihood function

Standard result:

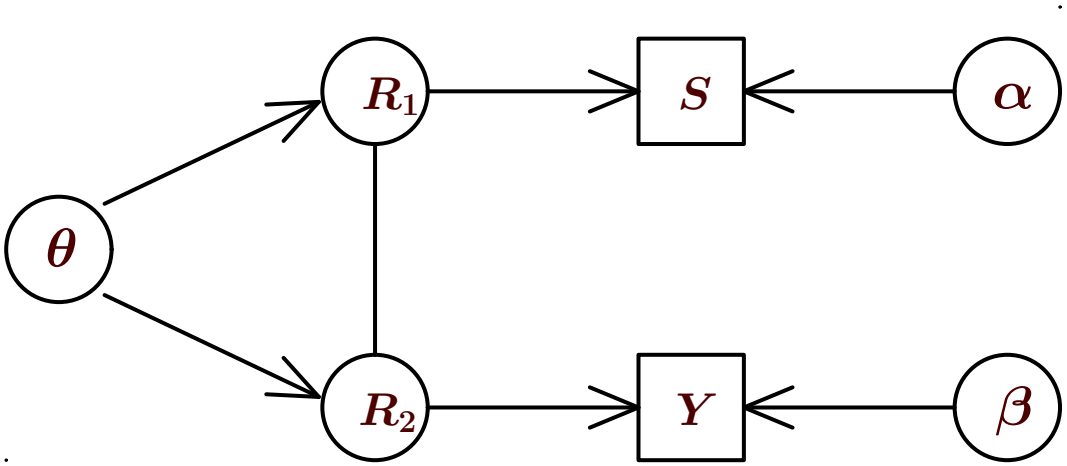
- $\log S|Y \sim \mathrm{N}(\mu_{S|Y}, \sigma^2_{S|Y})$
- $\mu_{S|Y} = \mu_S + \gamma' V(\phi)^{-1}(Y - \mu_Y)$
- $\sigma^2_{S|Y} = \sigma^2 - \gamma' V(\phi)^{-1}\gamma$

Likelihood contribution from ith subject:

- uncensored S_i:
  $[Y_i] \times [\log S_i|Y_i]$ (multivariate Gaussian pdf)
- censored S_i:
  $[Y_i] \times \{1 - \Phi((\log S_i - \mu_{S|Y_i})/\sigma_{S|Y})\}$ (pdf times tail probability)

Joint modelling

Random effects models



Formulation of random effects models

Latent stochastic process

Bivariate Gaussian process $\boldsymbol{R}(t) = \{\boldsymbol{R}_1(t), \boldsymbol{R}_2(t)\}$

- $\boldsymbol{R}_k(t) = \boldsymbol{D}_k(t)\boldsymbol{U}_k + \boldsymbol{W}_k(t)$
- $\{\boldsymbol{W}_1(t), \boldsymbol{W}_2(t)\}$: bivariate stationary Gaussian process
- $(\boldsymbol{U}_1, \boldsymbol{U}_2)$: multivariate Gaussian random effects

Bivariate process $\boldsymbol{R}(t)$ is realised independently between subjects

Measurement sub-model

$$\boldsymbol{Y_{ij} = \mu_i(t_{ij}) + R_{1i}(t_{ij}) + Z_{ij}}$$

- $\boldsymbol{Z_{ij} \sim \mathrm{N}(0, \tau^2)}$
- $\boldsymbol{\mu_i(t_{ij}) = X_{1i}(t_{ij})\beta_1}$

Hazard sub-model

$$\boldsymbol{h_i(t) = h_0(t)\mathcal{F}\{X_2(t)\beta_2 + R_{2i}(t)\}}$$

- $\boldsymbol{h_0(t)}$ **= non-parametric baseline hazard**
- $\boldsymbol{\eta_2(t) = X_{2i}(t) + R_{2i}(t)}$ **= linear predictor for hazard**
- **typical choice** $\boldsymbol{\mathcal{F}(\cdot) = \exp(\cdot)}$

Random effects models: the likelihood function

- conditional independence: $S \perp Y|R$
- standard Gaussian marginal: $[Y]]$,$L_1(\theta; Y)$
- Gaussian conditional: $[R|Y]$
- standard conditional: $[S|R]$, $L_2(\theta; S|R)$
- selection factorisation

$$\begin{aligned}[Y, S] &= \int [Y, S, R] dR \\ &= \int [Y][R|Y][S|R, Y] dR \\ &= [Y] \int [R|Y][S|R] dR\end{aligned}$$

$$L(\theta; Y, S) = L_1(\theta; Y) \times \mathrm{E}_{R|Y}[L_2(\theta; S|R)]$$

Evaluating the likelihood function

$$L(\theta; Y, S) = L_1(\theta; Y) \times \mathrm{E}_{R|Y}[L_2(\theta; S|R)]$$

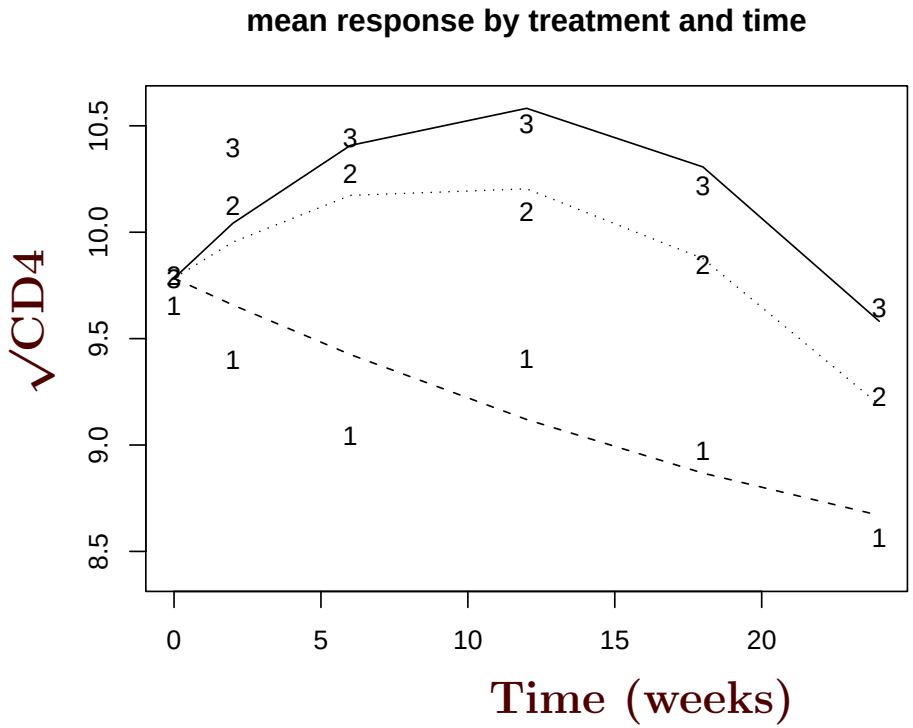
- R is infinite-dimensional, but
- non-parametric specification of $h_0(\cdot)$ implies we only need R at event-times
- Monte Carlo evaluation of expectation term
- explicit EM evaluation possible in useful special cases (eg Wulfsohn and Tsiatis, 1997)

Examples

- **AIDS data (transformation model)**
- **Schizophrenia data (random effects model)**
- **Heart surgery data (random effects model?)**

AIDS data

Mean square-root CD4



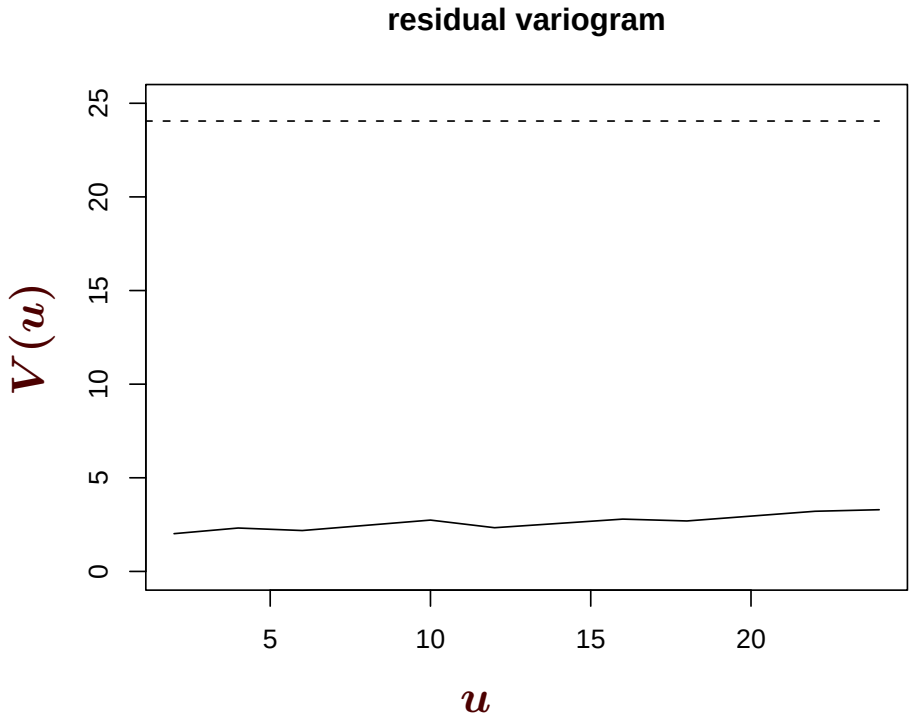
Fitted lines are quadratics with common intercept

Covariance structure

- **Variance matrix of OLS residuals from saturated treatments-by-times model**

```
17.94 18.26 18.53 18.80 18.47 18.01
18.26 22.86 21.02 21.32 20.94 20.57
18.53 21.02 23.95 22.47 22.02 21.76
18.80 21.32 22.47 25.01 22.60 22.54
18.47 20.94 22.02 22.60 24.67 22.20
18.01 20.57 21.76 22.54 22.20 24.64
```

- Variogram tells similar story: dominant source of variation is between subjects

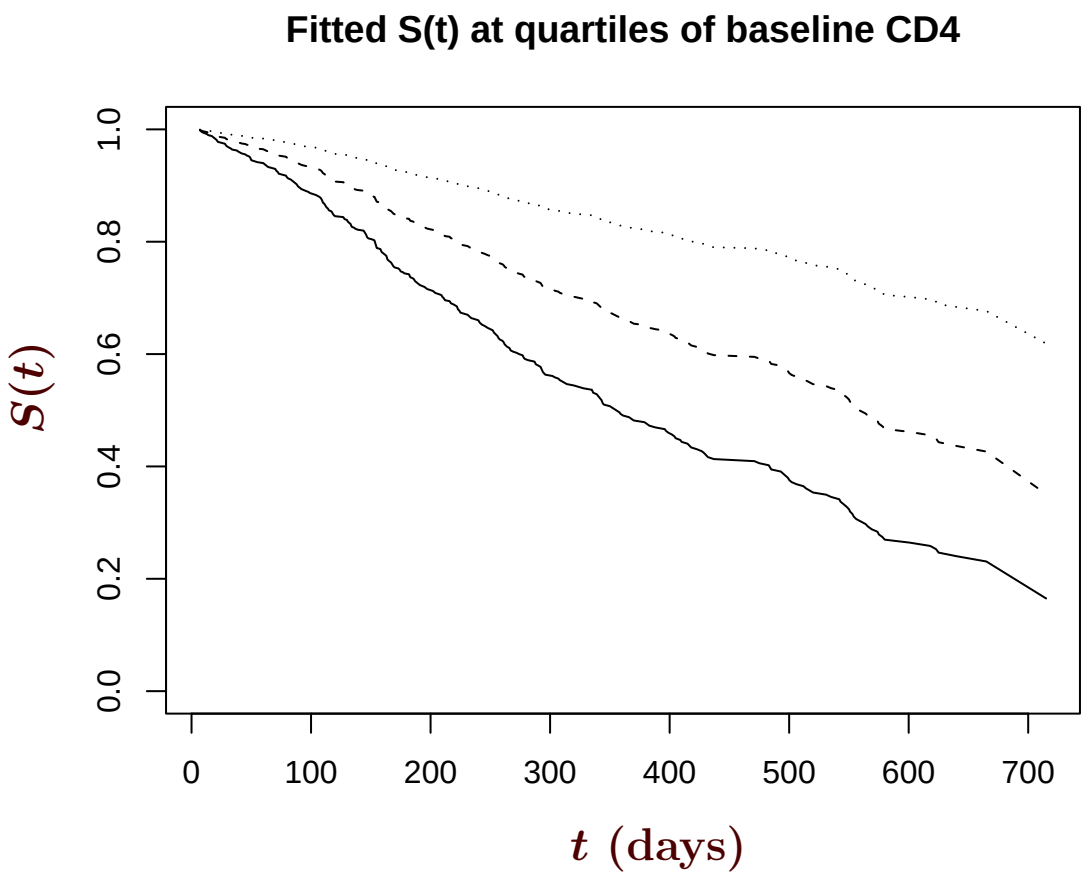


$R(t) =$ residual at time t

$V(u) = \frac{1}{2}\text{Var}\{R(t) - R(t-u)\}$

AIDS data

Baseline CD4 and time to progression/death



Model formulation

- exchangeable covariance structure for $\boldsymbol{Y}$
- non-parametric specification of $\mathrm{Cov}(\log \boldsymbol{S}, \boldsymbol{Y})$
- mean $\sqrt{\text{CD4}}$ quadratic in time within each treatment group
- linear covariate adjustments

Maximum likelihood estimation

- **Treatment contrasts:**

  $\theta_{2-1}, \theta_{3-1}$: contrasts in mean $\log S$

  $\theta^*_{2-1}, \theta^*_{3-1}$: contrasts in mean Y at 24 weeks

Scale	parameter	estimate	std. error	correlation
$\log S$	θ_{2-1}	0.351	0.129	
	θ_{3-1}	0.222	0.126	0.486
Y	θ^*_{2-1}	0.981	0.439	
	θ^*_{3-1}	1.003	0.362	0.423

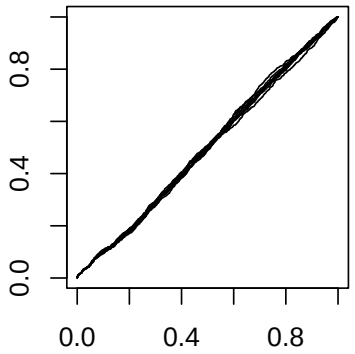
- **Covariance structure**

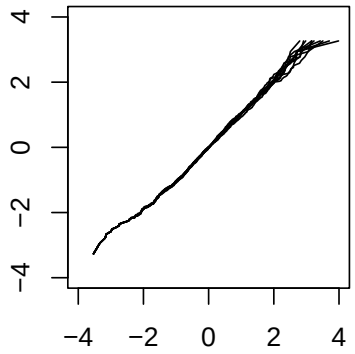
$$\mathrm{Var}(\boldsymbol{Y}) = \tau^2 \boldsymbol{I} + \nu^2 \boldsymbol{J}$$
$$\mathrm{Var}(\log \boldsymbol{S}) = \sigma^2$$
$$\mathrm{Corr}(\log \boldsymbol{S}, \boldsymbol{Y}_j) = \rho_j$$

Parameter	estimate	std. error
τ^2	2.432	0.057
ν^2	16.723	0.808
σ^2	1.653	0.145
ρ_1	0.428	0.038
ρ_2	0.468	0.035
ρ_3	0.467	0.035
ρ_4	0.457	0.036
ρ_5	0.520	0.035
ρ_6	0.509	0.038

Goodness-of-fit

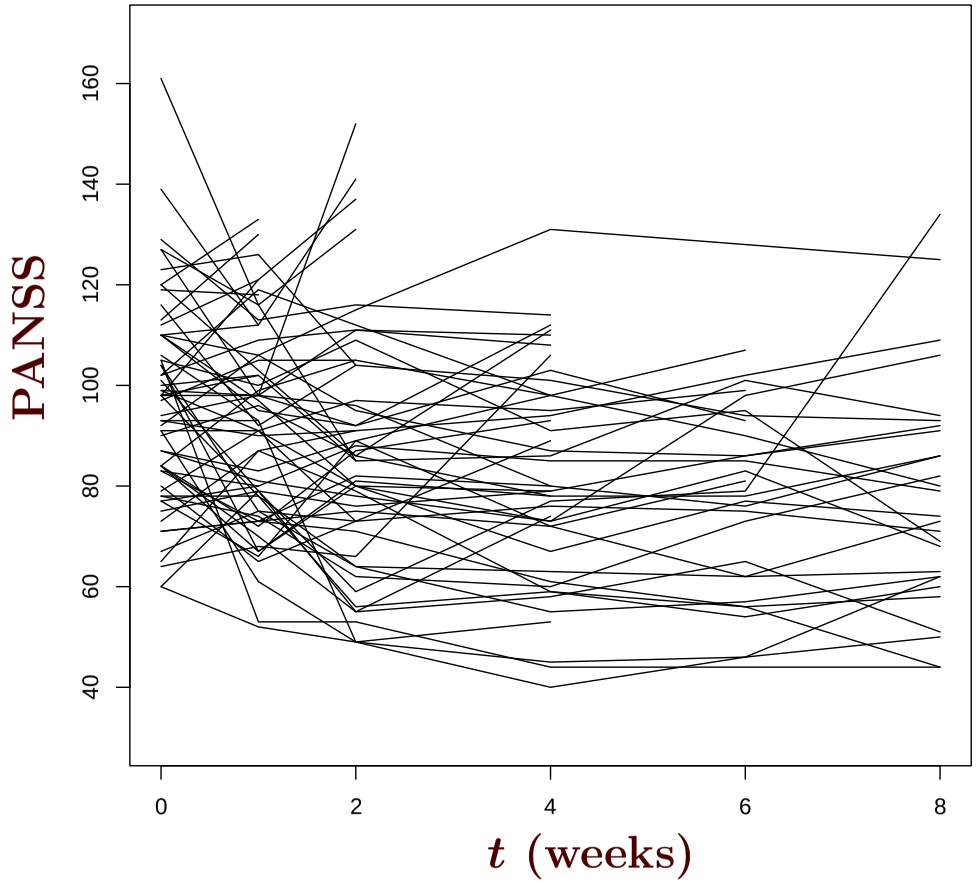
- Focus on conditional distribution of $\log S$ given Y
- Gaussian P-P and Q-Q plots with multiple imputation of censored $\log S$



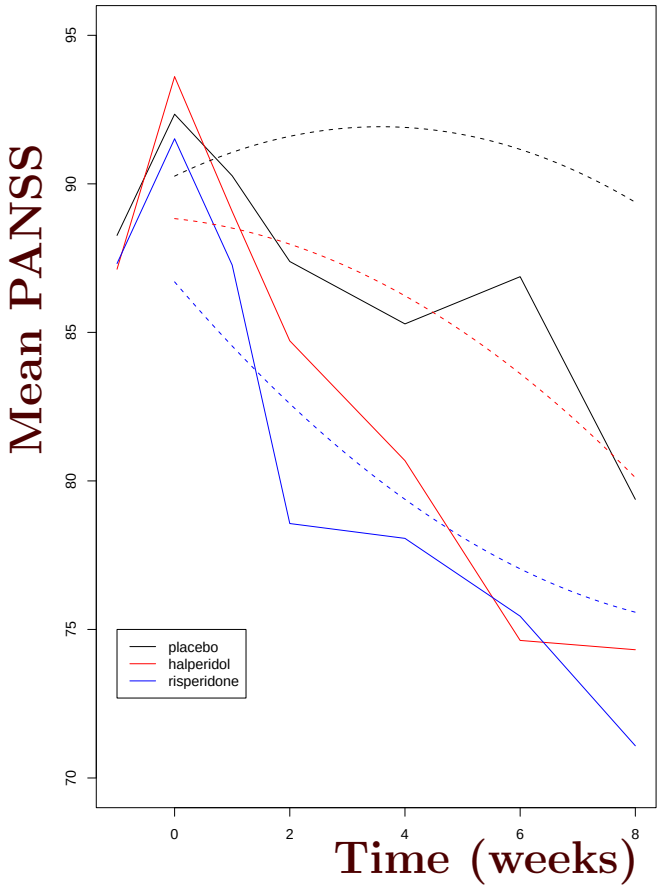


Schizophrenia data

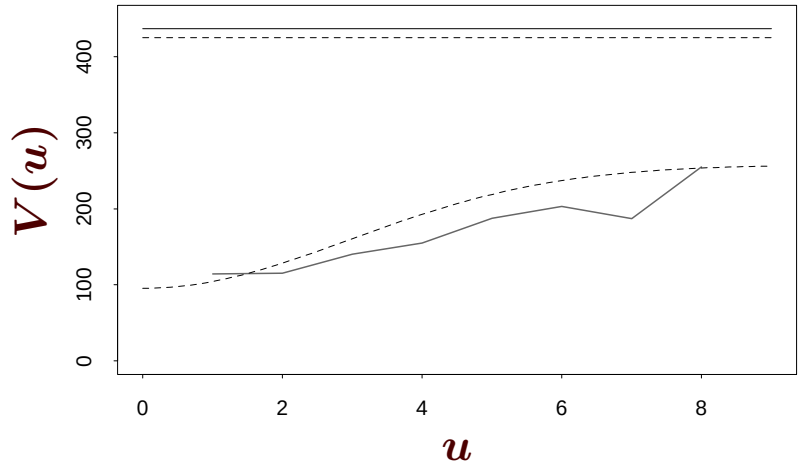
PANSS responses from haloperidol arm



Schizophrenia data
Mean response

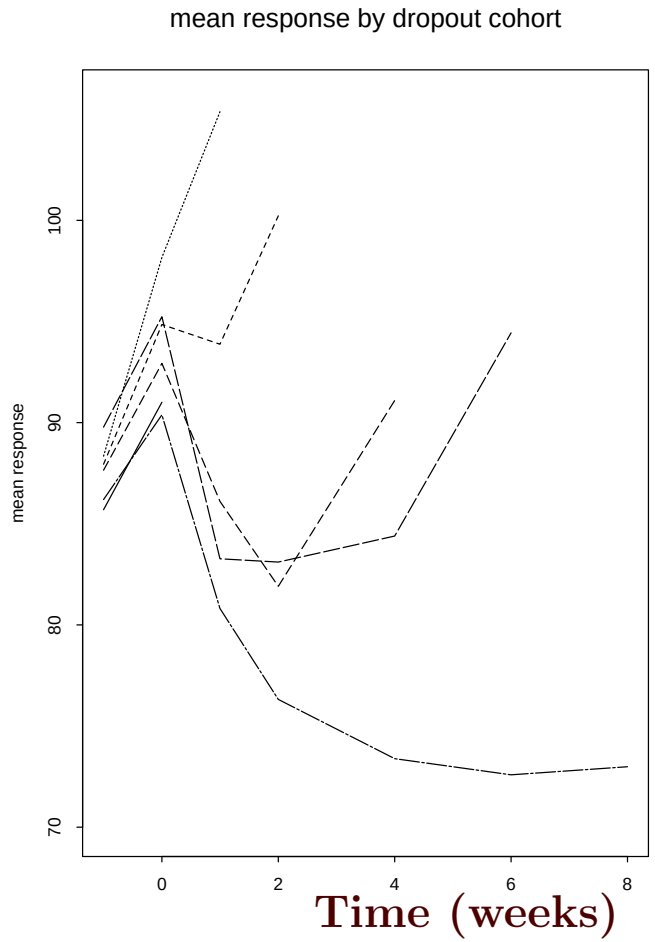


Empirical and fitted variograms



Apparently fits well, but hard to distinguish empirically from random intercept and slope model.

Mean response by dropout cohort



Model formulation

Measurement sub-model

For subject in treatment group k,

$$\mu_i(t) = \beta_{0k} + \beta_{1k}t + \beta_{2k}t^2$$

$$Y_{ij} = \mu_i(t_{ij}) + R_{1i}(t_{ij}) + Z_{ij}$$

Hazard sub-model

For subject in treatment group k,

$$h_i(t) = h_0(t)\exp\{\alpha_k + R_{2i}(t)\}$$

Latent process

Two choices for measurement process component

$$R_1(t) = U_1 + W_1(t)$$

$$R_1(t) = U_1 + U_2 t$$

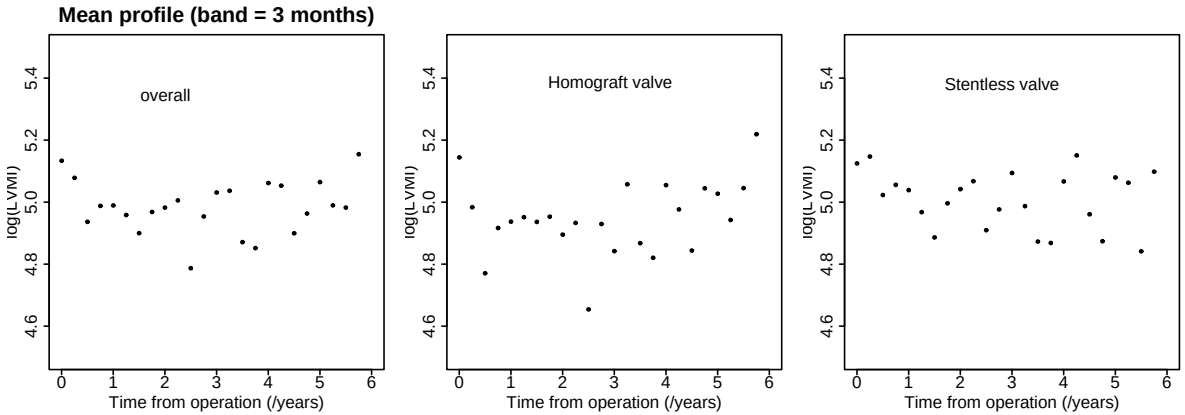
And for hazard process component

$$R_2(t) = \gamma_1 R_1(t)$$

$$R_2(t) = \gamma_1(U_1 + U_2 t) + \gamma_2 U_2$$

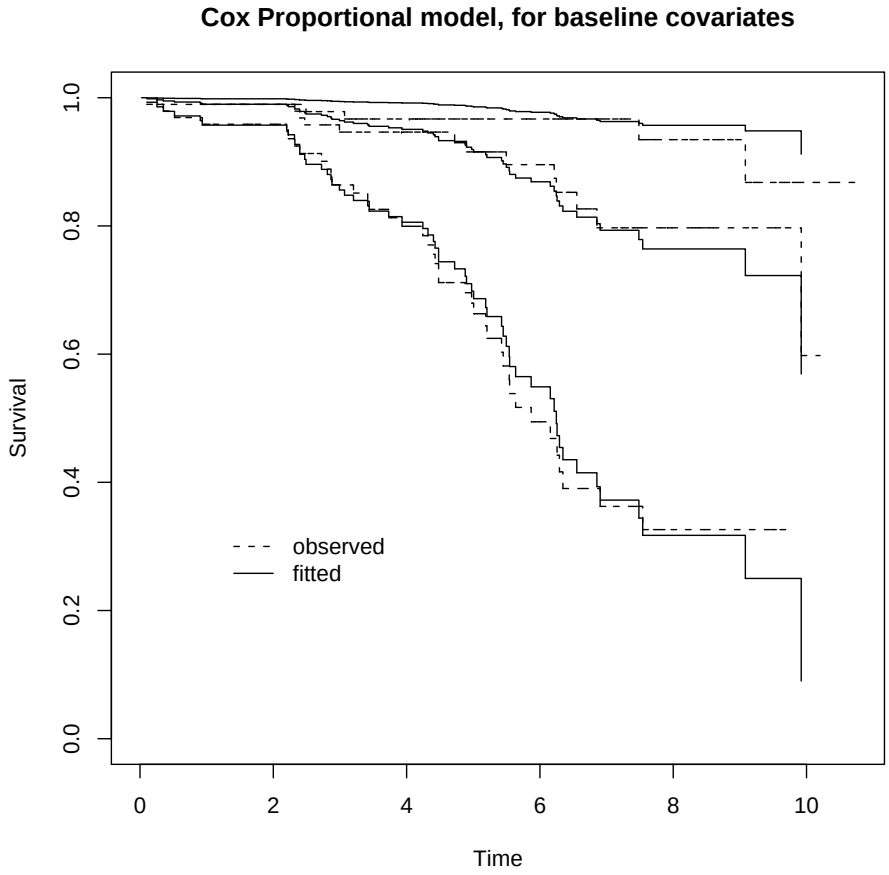
Heart surgery data

Mean log-LVMI response profiles



Heart surgery data

Survival curves adjusted for baseline covariates



Heart surgery data

Hypothesis

- subjects who regress after surgery (increasing LVMI) are at greater risk of heart attack

Exploratory analysis

- overall mean log(LVMI) decreases initially after surgery, then remains approximately constant
- but some patients regress (increasing LVMI)
- time-averaged log(LVMI) is associated with increased hazard

Random effects joint model for heart surgery data

$$Y_{ij} = \begin{cases} \mu(t_{ij}) + A_i + W_i(t_{ij}) + Z_j & : t \leq \tau \\ \mu(t_{ij}) + (A_i + \{B_i(t-\tau)\} + W_i(t_{ij}) + Z_j & : t > \tau \end{cases}$$

$$h_i(t) = h_0(t) \exp(x_i'\beta + B_i)$$

- surgery has immediate beneficial effect on all patients
- patient outcomes diverge after time τ post-surgery
- hazard for survival depends on slope of LVMI after time τ

Conclusions

- Choice of model/method should relate to scientific purpose.
- Simple models/methods are useful when exploring a range of modelling options, for example to select from amongst potential covariates.
- Complex models/methods are useful when seeking to understand subject-level stochastic variation.
- Flat likelihoods are common: different models may fit the data almost equally well.
- We need an `R` library.

References

Andersen, P.K., Borgan, Ø., Gill, R.D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. New York: Springer-Verlag.

Cox, D.R. (1972). Regression models and life tables (with Discussion). *Journal of the Royal Statistical Society* B, **34**, 187–220.

Cox, D.R. (1999). Some remarks on failure-times, surrogate markers, degradation, wear and the quality of life. *Lifetime Data Analysis*, **5**, 307–14.

Diggle, P.J., Heagerty, P., Liang, K-Y and Zeger, S.L. (2002). *Analysis of Longitudinal Data (second edition)*. Oxford : Oxford University Press.

Faucett, C.L. and Thomas, D.C. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Statistics in Medicine*, **15**, 1663–1686.

Finkelstein, D.M. and Schoenfeld, D.A. (1999). Combining mortality and longitudinal measures in clinical trials. *Statistics in Medicine*, **18**, 1341–1354.

Henderson, R., Diggle, P. and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, **1**, 465–80.

Henderson, R., Diggle, P. and Dobson, A. (2002). Identification and efficacy of longitudinal markers for survival. *Biostatistics*, **3**, 33–50.

Hogan, J.W. and Laird, N.M. (1997a). Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine* **16**, 239–257.

Hogan, J.W. and Laird, N.M. (1997b). Model-based approaches to analysing incomplete longitudinal and failure time data. *Statistics in Medicine* **16**, 259–272.

Little, R.J.A. (1995). Modelling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, **90**, 1112–21.

Pawitan, Y. and Self, S. (1993). Modelling disease marker processes in AIDS. *Journal of the American Statisical Association* **88**, 719–726.

Taylor, J.M.G., Cumberland, W.G. and Sy, J.P. (1994). A stochastic model for analysis of longitudinal AIDS data. *Journal of the American Statistical Association* **89**, 727–736

Tsiatis, A.A., Degruttola, V. and Wulsohn, M.S. (1995). Modelling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association* **90**, 27–37.

Wulfsohn, M.S. and Tsiatis, A.A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, **53**, 330–9.

Xu, J. and Zeger, S.L. (2001). Joint analysis of longitudinal data comprising repeated measures and times to events. *Applied Statistics*, **50**, 375–387.