A Short Course in Longitudinal Data Analysis

Peter J Diggle
Nicola Reeve, Michelle Stanton

(School of Health and Medicine, Lancaster University)

Lancaster, June 2011

Timetable

Day 1

9.00	**Registration**	
9.30	**Lecture 1**	**Motivating examples, exploratory analysis**
11.00	**BREAK**	
11.30	**Lab 1**	**Introduction to `R`**
12.30	**LUNCH**	
13.30	**Lecture 2**	**Linear modelling of repeated measurements**
15.00	**BREAK**	
15.30	**Lab 2**	**Exploring longitudunal data**
17.00	**CLOSE**	

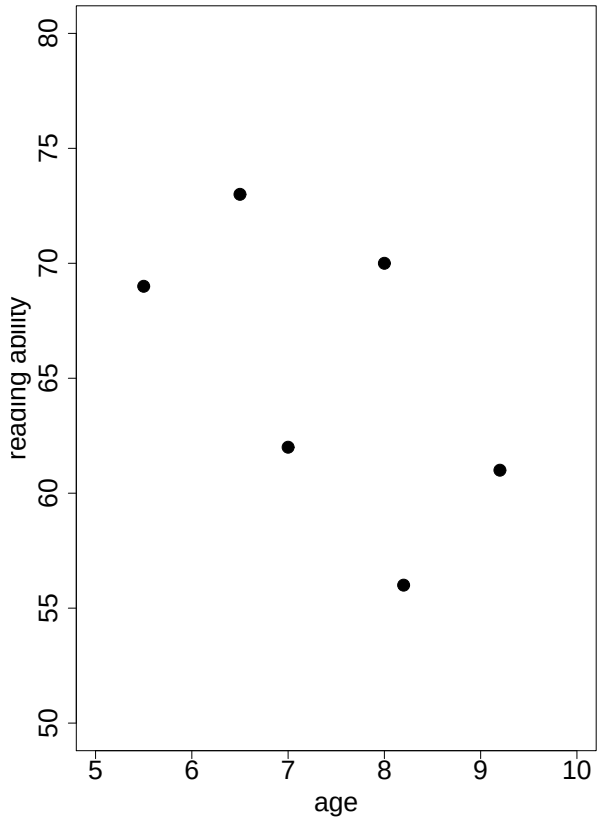
Day 2

9.00	**Lecture 3**	**Generalized linear models (GLM's)**
10.00	**Lab 3**	**The `nlme` package**
11.30	**BREAK**	
12.00	**Lecture 4**	**Joint modelling**
13.00	**Lunch**	
14.00	**Lab 4**	**Marginal and random effects GLM's**
16.00	**CLOSE**	

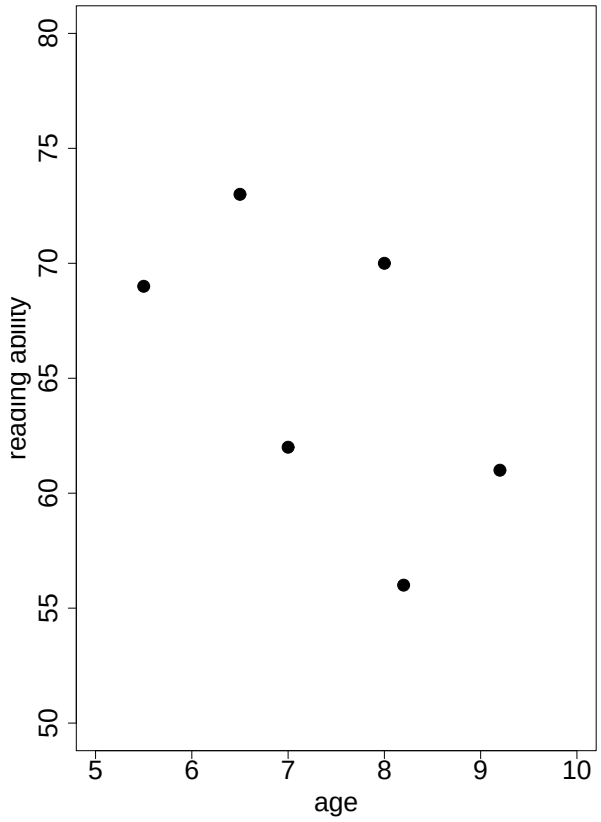
Lecture 1

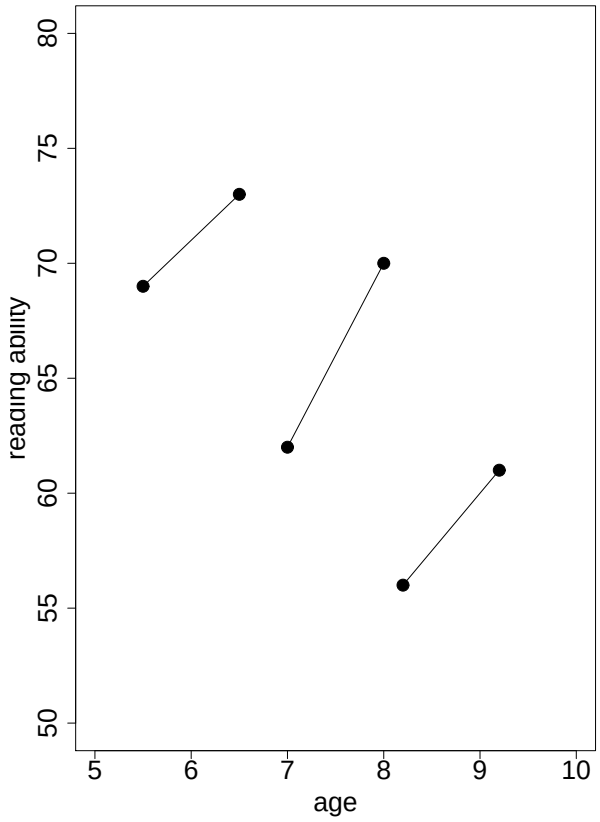
- examples
- scientific objectives
- why longitudinal data are correlated and why this matters
- balanced and unbalanced data
- tabular and graphical summaries
- exploring mean response profiles
- exploring correlation structure

Example 1. Reading ability and age

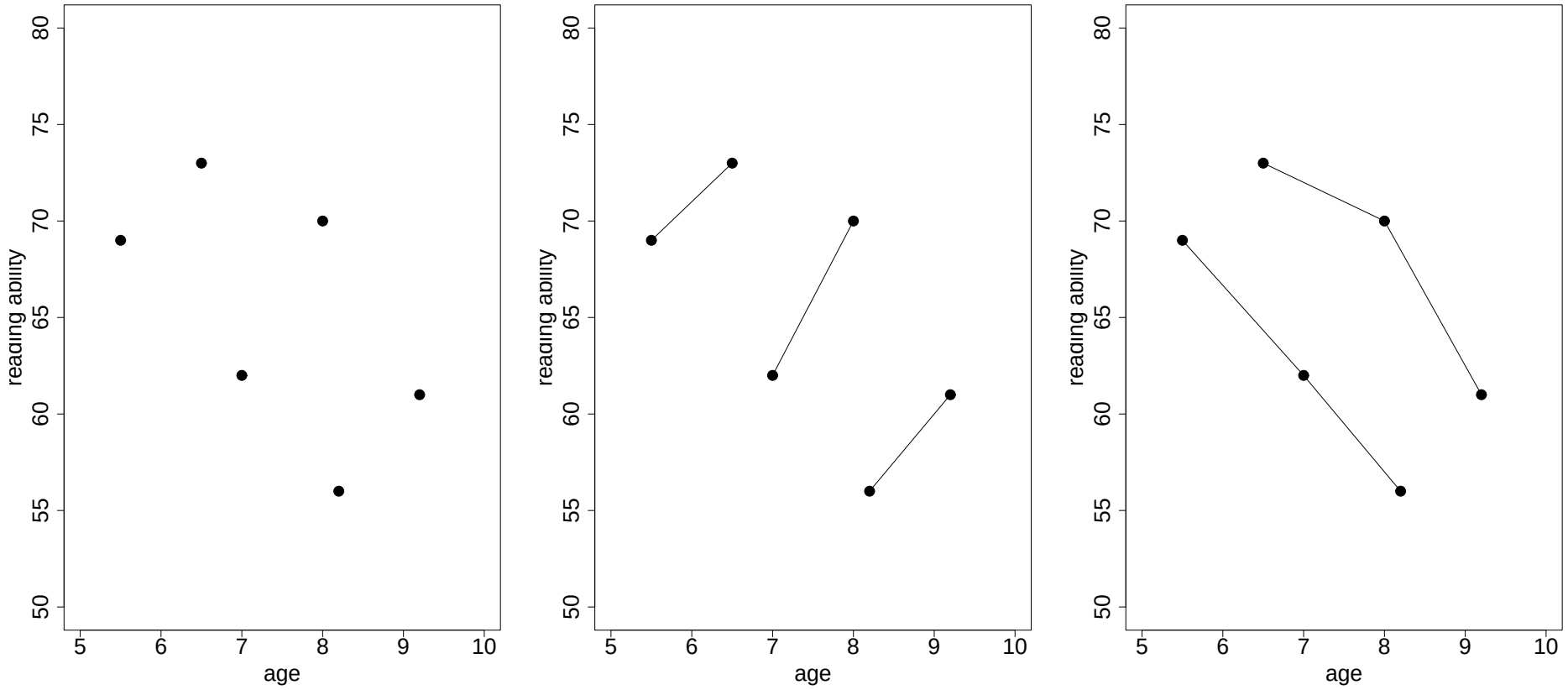


Example 1. Reading ability and age





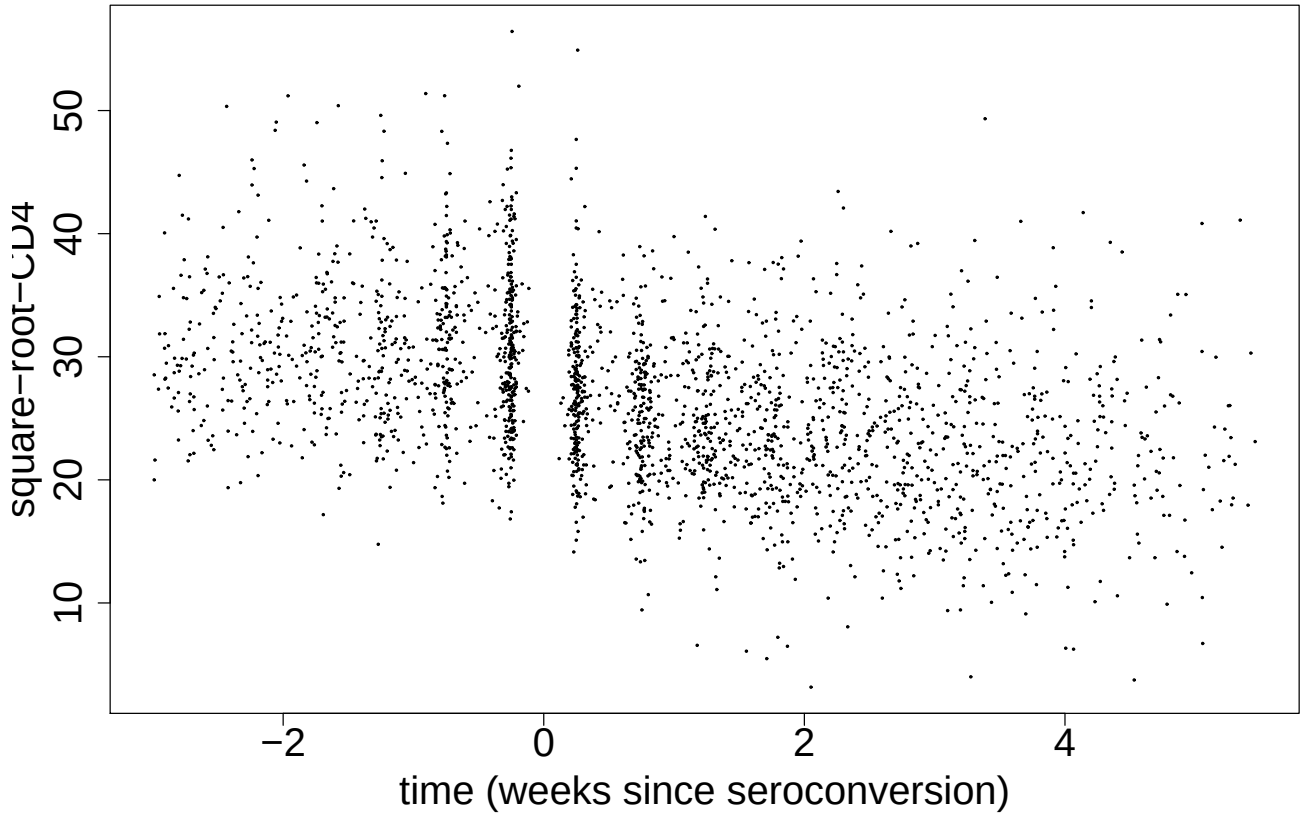
Example 1. Reading ability and age



Longitudinal *designs* enable us to distinguish cross-sectional and longitudinal *effects*.

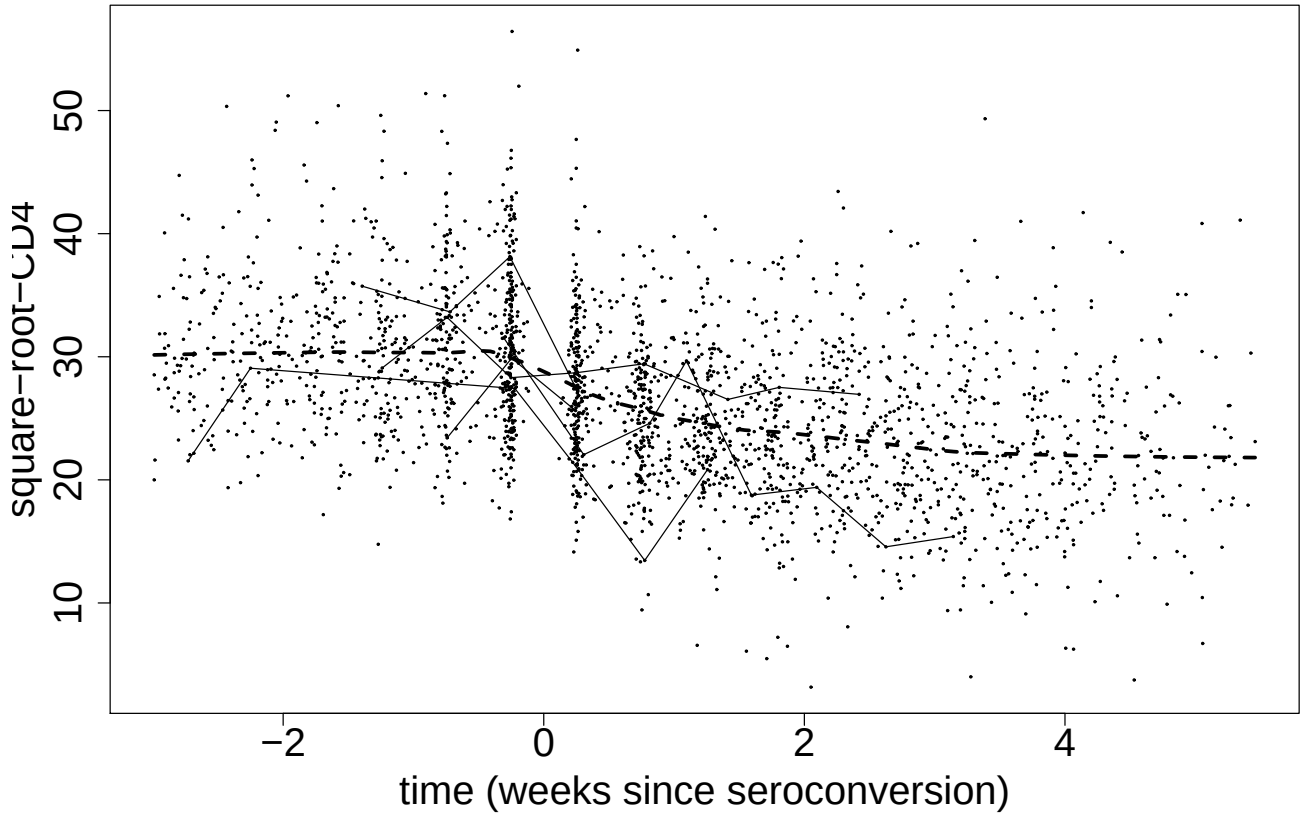
Example 2. CD4+ cell numbers

Cohort of 369 HIV seroconverters, CD4+ cell-count measured at approximately six-month intervals, variable number of measurements per subject.



Example 2. CD4+ cell numbers

Cohort of 369 HIV seroconverters, CD4+ cell-count measured at approximately six-month intervals, variable number of measurements per subject.

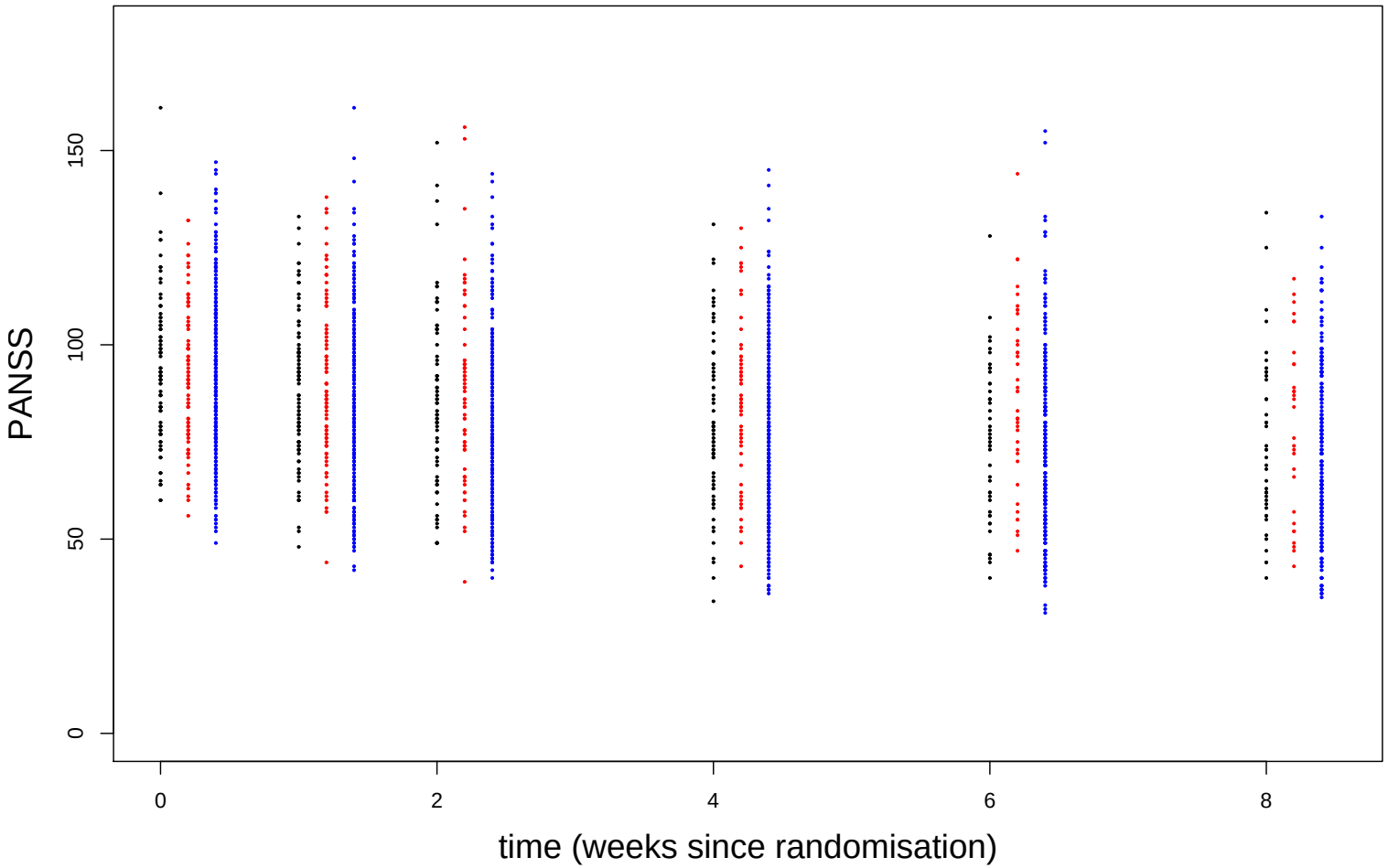


Example 3. Schizophrenia trial

- randomised clinical trial of drug therapies
- three treatments:
  - haloperidol (standard)
  - placebo
  - risperidone (novel)
- dropout due to "inadequate response to treatment"

Treatment	Number of non-dropouts at week					
	0	1	2	4	6	8
haloperidol	85	83	74	64	46	41
placebo	88	86	70	56	40	29
risperidone	345	340	307	276	229	199
total	518	509	451	396	315	269

Schizophrenia trial data (PANSS)



Scientific Objectives

Pragmatic philosophy: method of analysis should take account of the scientific goals of the study.

All models are wrong, but some models are useful

G.E.P. Box

- **scientific understanding or empirical description?**
- **individual-level or population-level focus?**
- **mean response or variation about the mean?**

Example 5. Smoking and health

- public health perspective – how would smoking reduction policies/programmes affect the health of the community?
- clinical perspective – how would smoking reduction affect the health of my patient?

Correlation and why it matters

- **different measurements on the same subject are typically correlated**
- **and this must be recognised in the inferential process.**

Estimating the mean of a time series

$$Y_1, Y_2, ..., Y_t, ..., Y_n \qquad Y_t \sim \mathrm{N}(\mu, \sigma^2)$$

Classical result from elementary statistical theory:

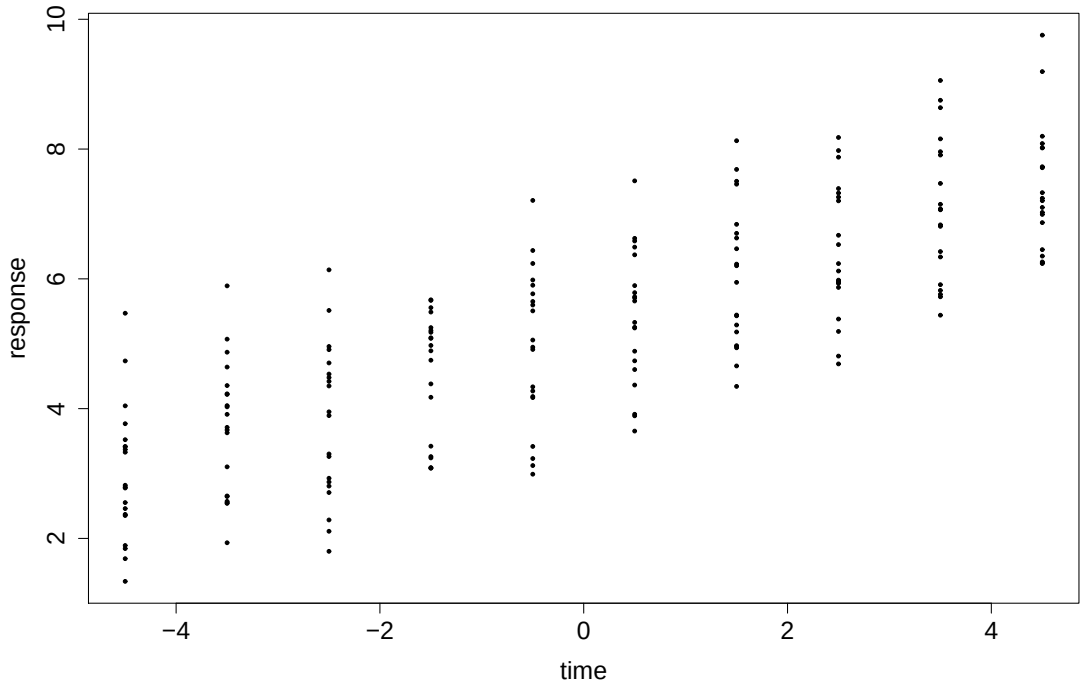
$$\bar{Y} \pm 2\sqrt{\sigma^2/n}$$

But if Y_t is a time series:

- $\mathrm{E}[\bar{Y}] = \mu$
- $\mathrm{Var}\{\bar{Y}\} = (\sigma^2/n) \times \{1 + n^{-1} \sum_{u \neq t} \mathrm{Corr}(Y_t, Y_u)\}$

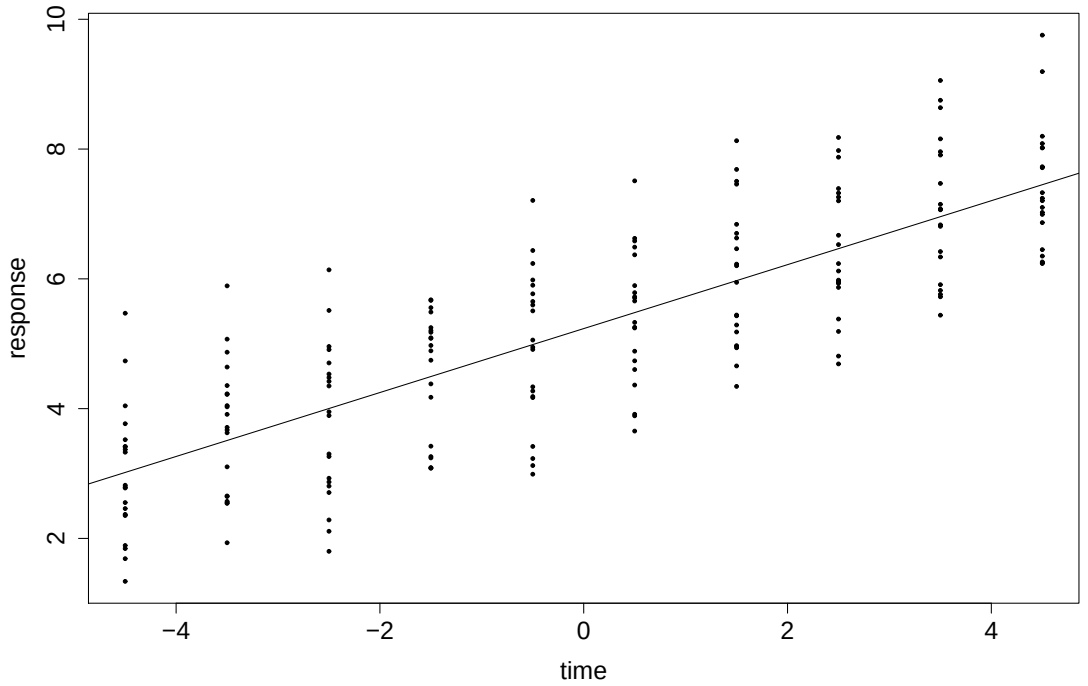
Correlation may or may not hurt you

$$Y_{it} = \alpha + \beta(t - \bar{t}) + Z_{it} \qquad i = 1, ..., m \quad t = 1, ..., n$$



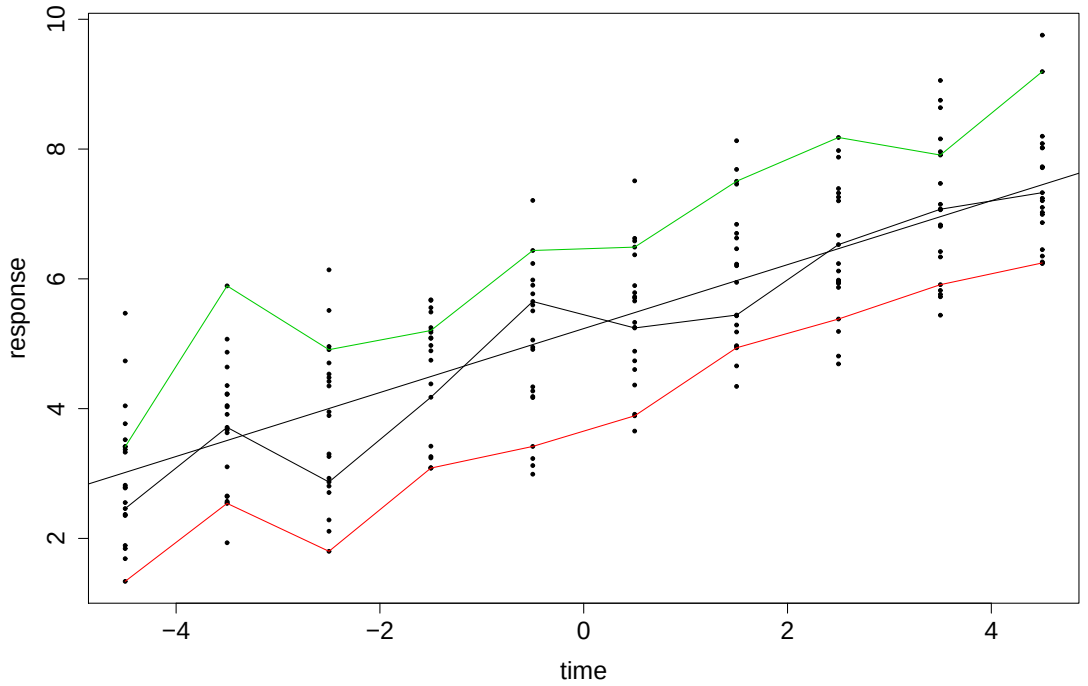
Correlation may or may not hurt you

$$Y_{it} = \alpha + \beta(t - \bar{t}) + Z_{it} \qquad i = 1, ..., m \quad t = 1, ..., n$$



Correlation may or may not hurt you

$$Y_{it} = \alpha + \beta(t - \bar{t}) + Z_{it} \qquad i = 1, ..., m \quad t = 1, ..., n$$



Correlation may or may not hurt you

$$Y_{it} = \alpha + \beta(t - \bar{t}) + Z_{it} \qquad i = 1, ..., m \quad t = 1, ..., n$$

Parameter estimates and standard errors:

	ignoring correlation		recognising correlation	
	estimate	standard error	estimate	standard error
α	5.234	0.074	5.234	0.202
β	0.493	0.026	0.493	0.011

Balanced and unbalanced designs

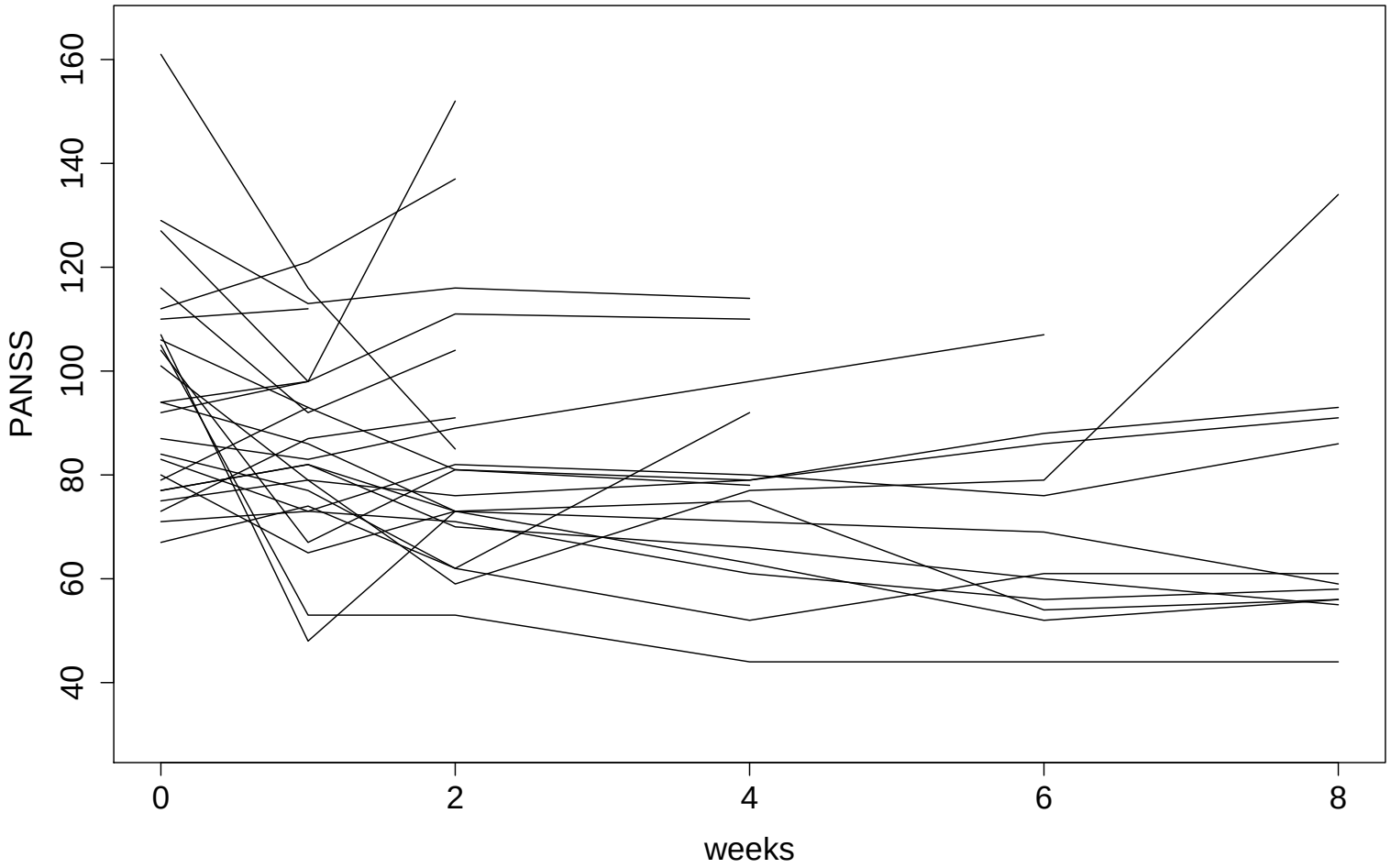
$$\begin{aligned} Y_{ij} = &\quad j^{th} \text{ measurement on } i^{th} \text{ subject} \\ t_{ij} = &\quad \text{time at which } Y_{ij} \text{ is measured} \end{aligned}$$

- balanced design: $t_{ij} = t_j$ for all subjects i
- a balanced design may generate unbalanced data

Missing values

- **dropout**
- **intermittent missing values**
- **loss-to-follow-up**

Random sample of PANSS response profiles



Tabular summary

PANSS treatment group 1 (standard drug)

Week	Mean	Variance	Correlation					
0	93.61	214.69	1.00	0.46	0.44	0.49	0.45	0.41
1	89.07	272.46	0.46	1.00	0.71	0.59	0.65	0.51
2	84.72	327.50	0.44	0.71	1.00	0.81	0.77	0.54
4	80.68	358.30	0.49	0.59	0.81	1.00	0.88	0.72
6	74.63	376.99	0.45	0.65	0.77	0.88	1.00	0.84
8	74.32	476.02	0.41	0.51	0.54	0.72	0.84	1.00

More than one treatment?

- separate tables for each treatment group
- look for similarities and differences

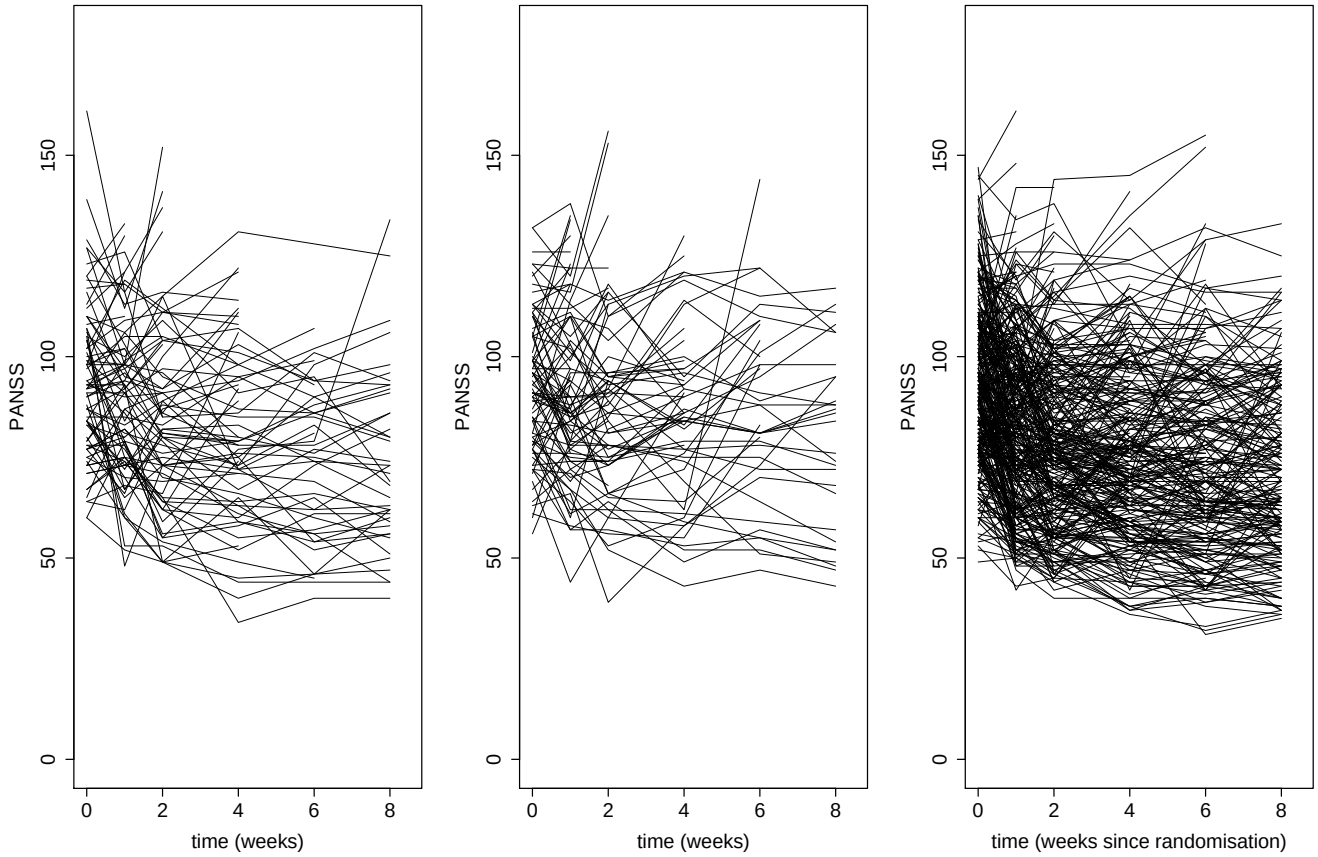
Covariates?

- use residuals from working model fitted by ordinary least squares

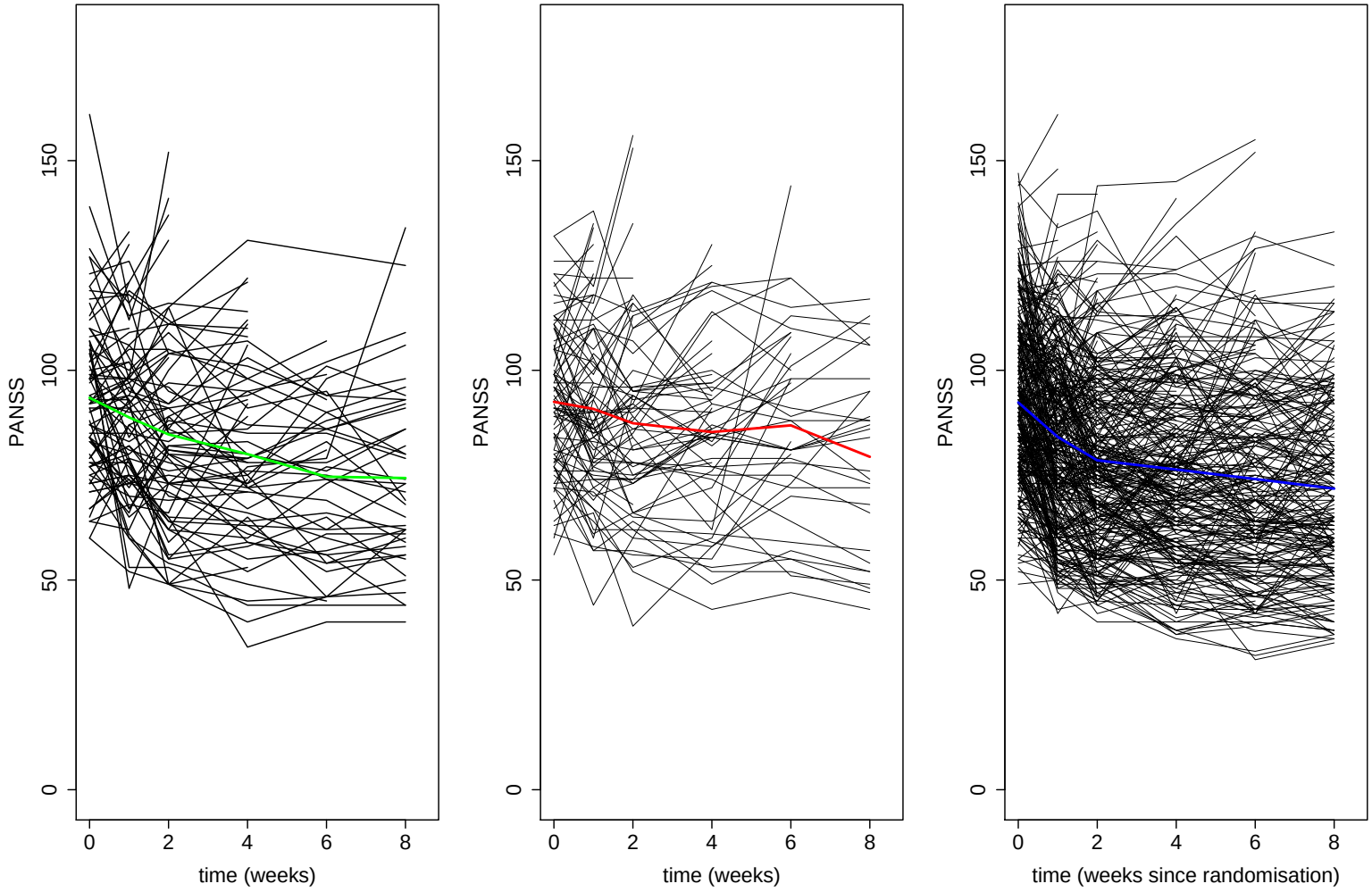
Graphical summary

- spaghetti plots
- mean response profiles
- non-parametric smoothing
- pairwise scatterplots
- variograms

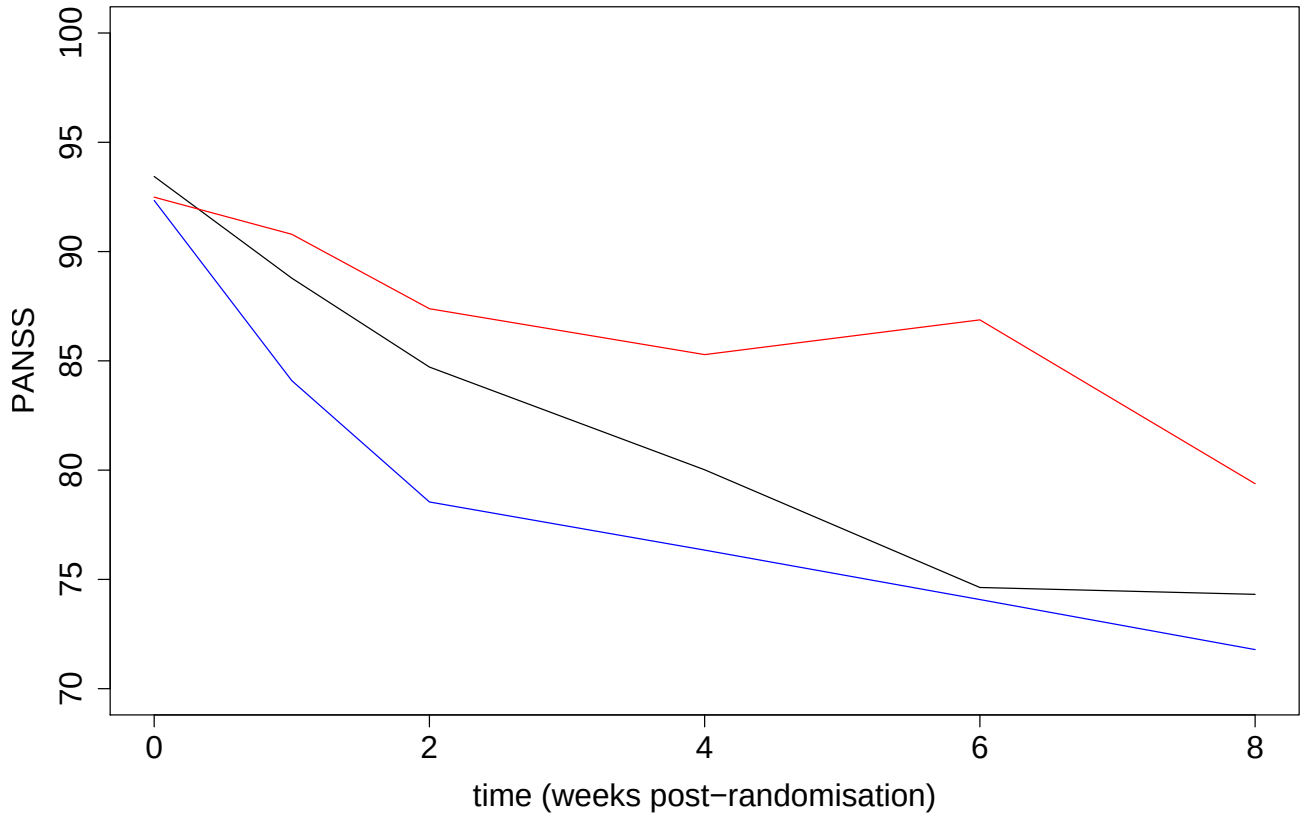
A spaghetti plot



A slightly better spaghetti plot



A set of mean response profiles



Kernel smoothing

- useful for unbalanced data
- simplest version is mean response within a moving time-window

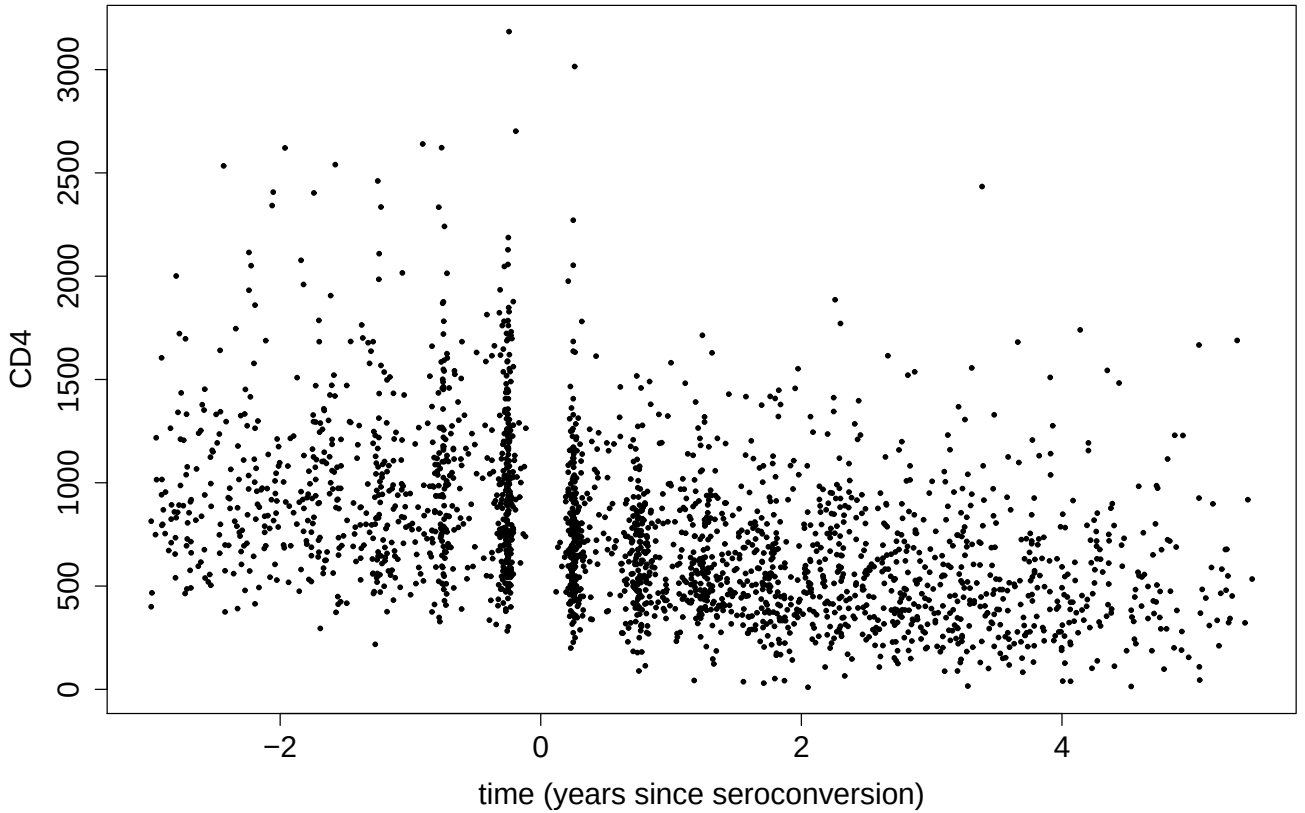
$$\hat{\mu}(t) = \text{average}(y_{ij} : |t_{ij} - t| < h/2)$$

- more sophisticated version:
  - kernel function $k(\cdot)$ (symmetric pdf)
  - band-width h

$$\hat{\mu}(t) = \sum y_{ij} k\{(t_{ij} - t)/h\} / \sum k\{(t_{ij} - t)/h\}$$

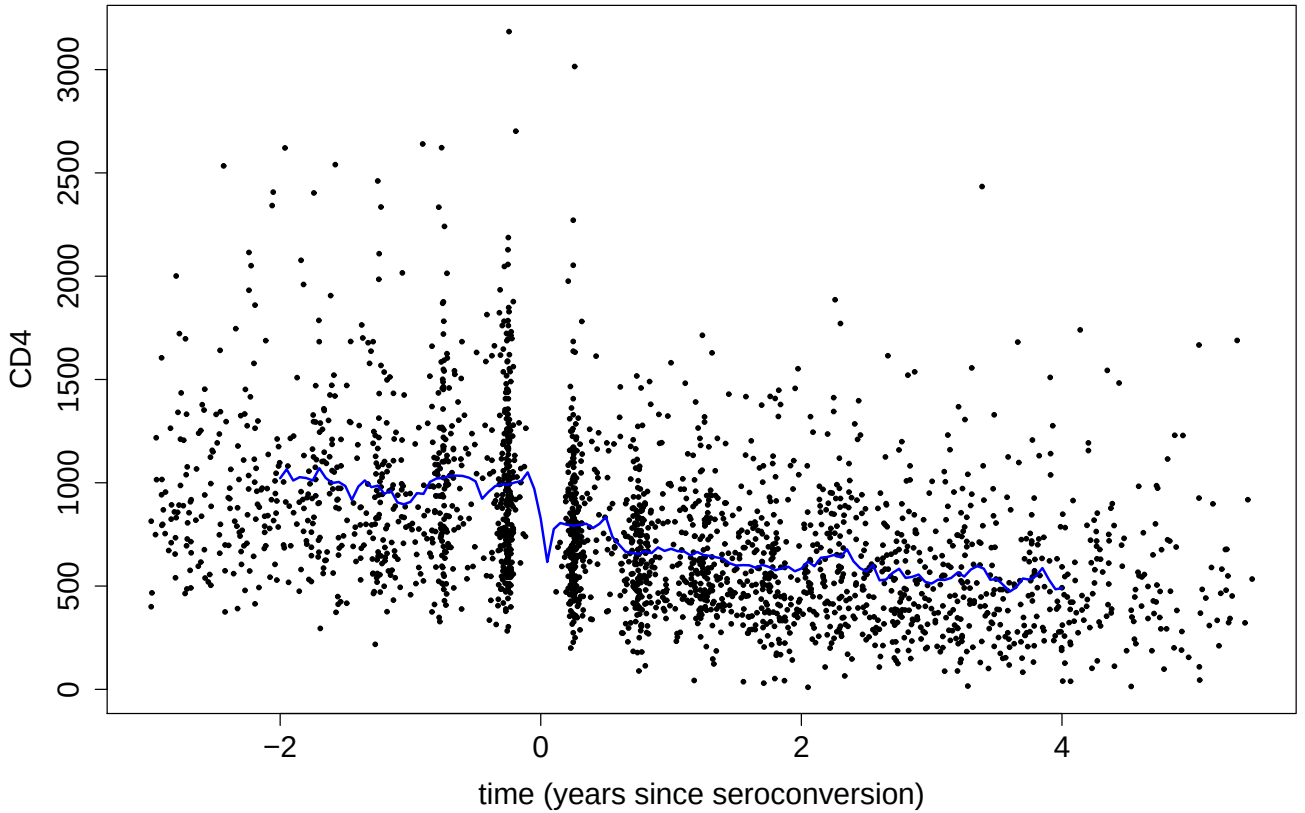
Smoothing the CD4 data

Data



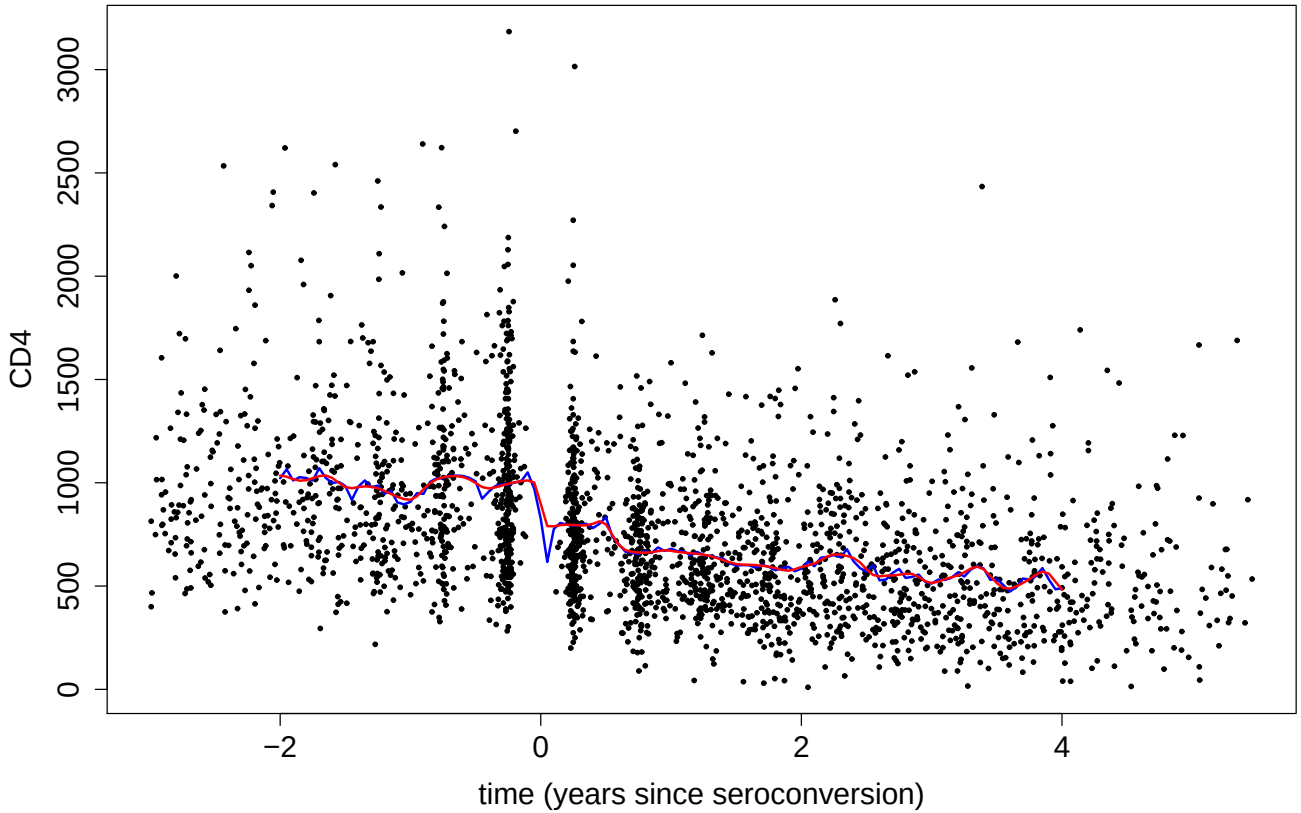
Smoothing the CD4 data

Data, uniform kernel



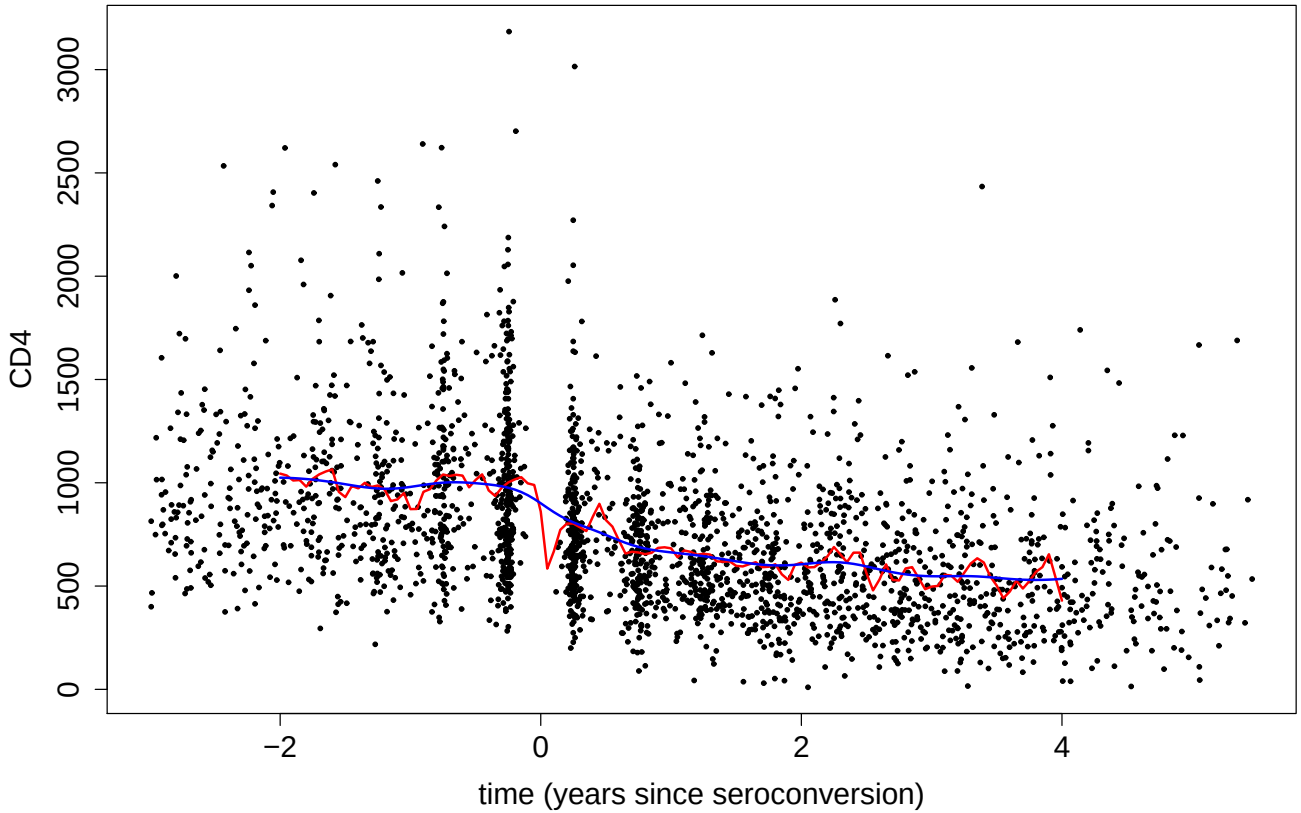
Smoothing the CD4 data

Data, uniform and Gaussian kernels



Smoothing the CD4 data

Data, Gaussian kernels with small and large band-widths



The variogram

The variogram of a stochastic process $Y(t)$ is

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

- well-defined for stationary and some non-stationary processes
- for stationary processes,

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

- easier to estimate $V(u)$ than $\rho(u)$ when data are unbalanced

Estimating the variogram

r_{ij} = residual from preliminary model for mean response

- Define

$$v_{ijk\ell} = \frac{1}{2}(r_{ij} - r_{k\ell})^2$$

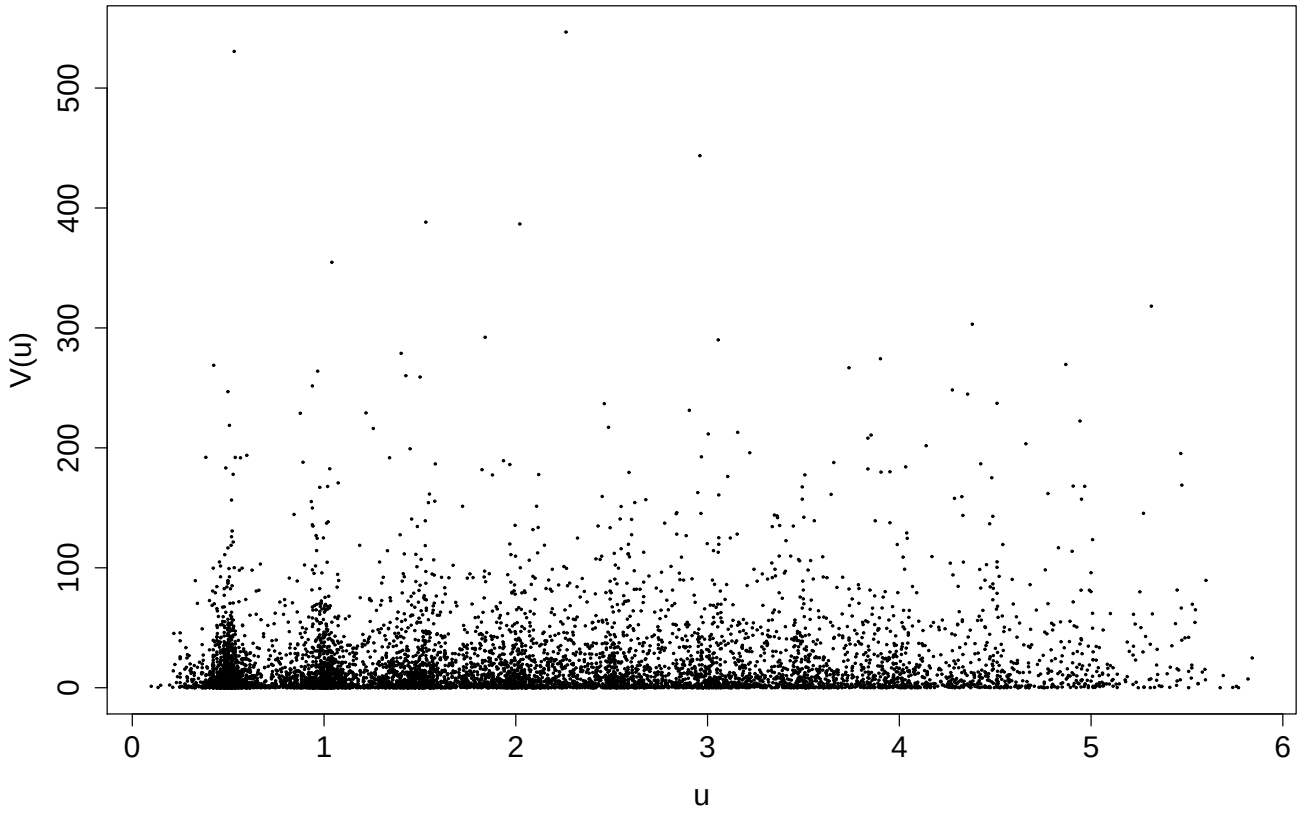
- Estimate

$\hat{V}(u)$ = average of all quantities $v_{iji\ell}$ such that $|t_{ij} - t_{i\ell}| \simeq u$

- Estimate of process variance

$\hat{\sigma}^2$ = average of all quantities $v_{ijk\ell}$ such that $i \neq k$.

Example 2. Square-root CD4+ cell numbers

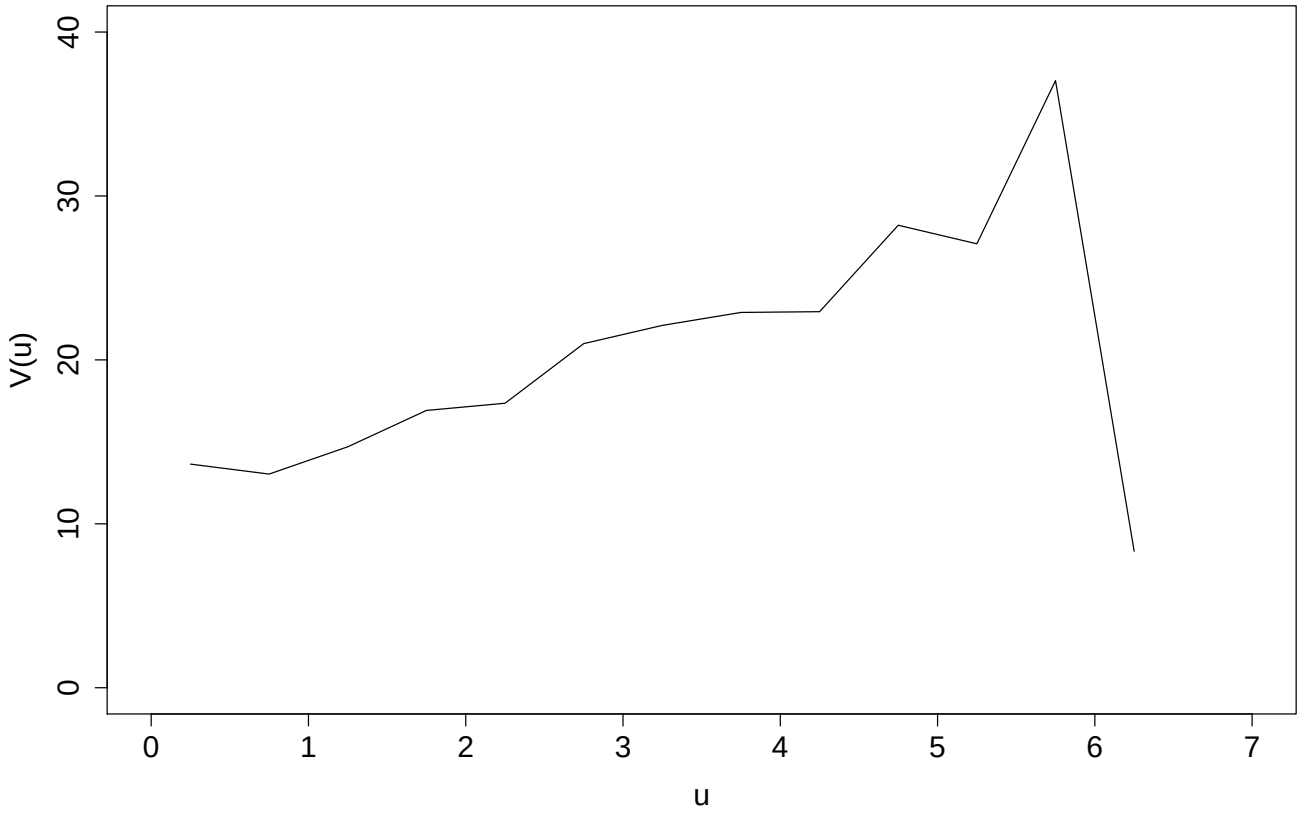


Very large sampling fluctuations hide the information

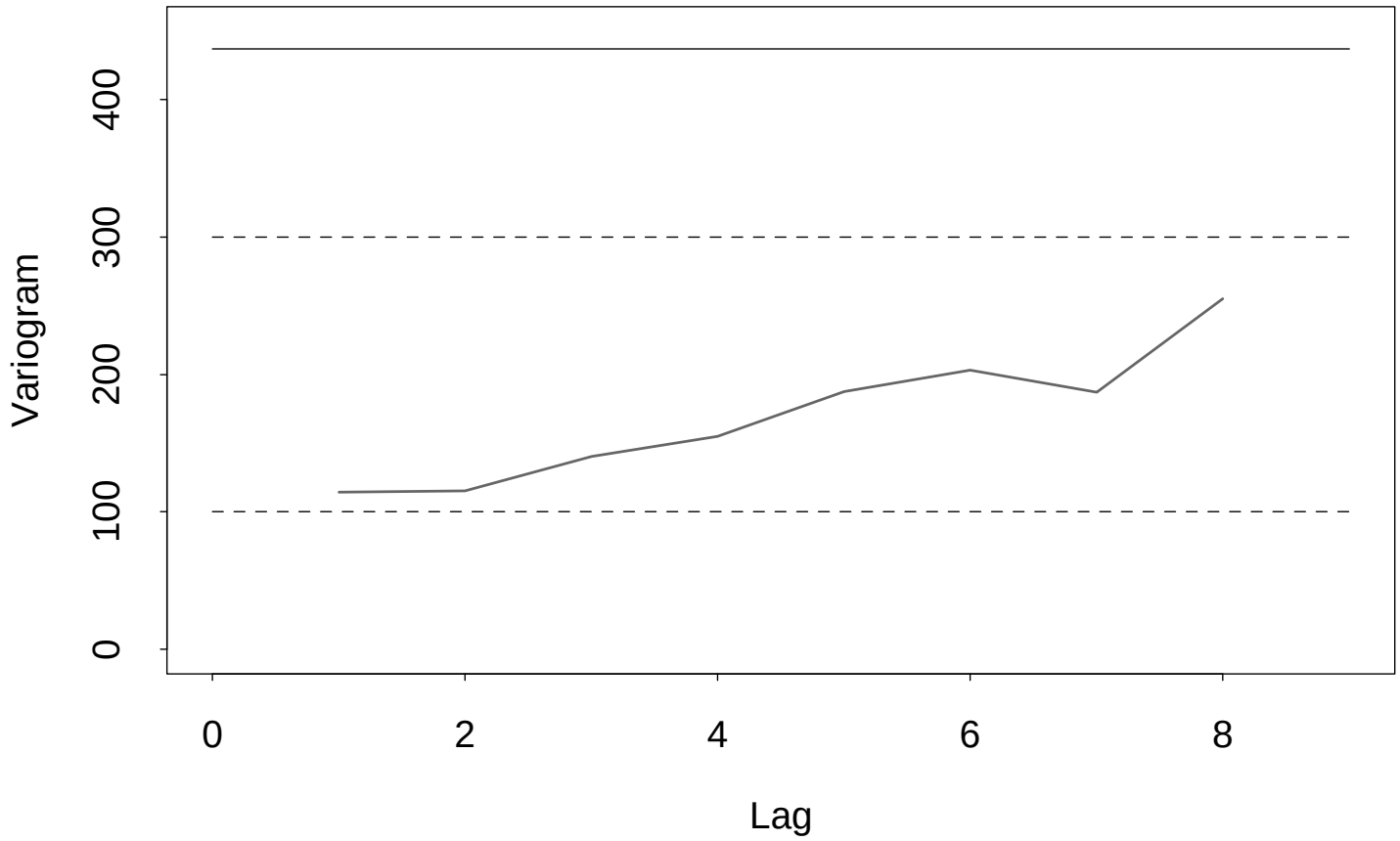
Smoothing the empirical variogram

- For irregularly spaced data:
  - group time-differences u into bands
  - take averages of corresponding v_{ijil}
- For data from a balanced design, usually no need to average over bands of values for u

Example 2. CD4+ cell numbers



Example 3. schizophrenia trial



Sampling distribution of the empirical variogram

- Gaussian distribution
- balanced data
- s subjects in each of p experimental groups
- n measurement times per subject
- mean responses estimated by ordinary least squares from saturated treatments-by-times model

Properties of $\hat{V}(u)$

- Marginal distribution of empirical variogram ordinates is

$$v_{ijil} \sim \{(s-1)/s\} V(u_{ijil}) \chi^2_1$$

- $\{s/(s-1)\}\hat{V}(u)$ is unbiased for $V(u)$
- expression available for covariance between any two quantities v_{ijil}
- hence can compute variance of $\hat{V}(u)$
- typically, $\mathrm{Var}\{\hat{V}(u)\}$ increases (sharply) as u increases

Where does the correlation come from?

- differences between subjects
- variation over time within subjects
- measurement error

```
data=read.table("CD4.data",header=T)
data[1:3,]
time=data$time
CD4=data$CD4
plot(time,CD4,pch=19,cex=0.25)
id=data$id
uid=unique(id)
for (i in 1:10) {
   take=(id==uid[i])
   lines(time[take],CD4[take],col=i,lwd=2)
   }
```

Lecture 2

- The general linear model with correlated residuals
- parametric models for the covariance structure
- the clever ostrich (why ordinary least squares may not be a silly thing to do)
- weighted least squares as maximum likelihood under Gaussian assumptions
- missing values and dropouts

General linear model, correlated residuals

$$
\begin{aligned}
E(Y_{ij}) &= x_{ij1}\beta_1 + \ldots + x_{ijp}\beta_p \\
Y_i &= X_i\beta + \epsilon_i \\
Y &= X\beta + \epsilon
\end{aligned}
$$

- measurements from different subjects independent
- measurements from same subject typically correlated.

Parametric models for covariance structure

Three sources of random variation in a typical set of longitudinal data:

- **Random effects (variation between subjects)**
  - **characteristics of individual subjects**
  - **for example, intrinsically high or low responders**
  - **influence extends to all measurements on the subject in question.**

Parametric models for covariance structure

Three sources of random variation in a typical set of longitudinal data:

- **Random effects**
- **Serial correlation (variation over time within subjects)**
  - **measurements taken close together in time typically more strongly correlated than those taken further apart in time**
  - **on a sufficiently small time-scale, this kind of structure is almost inevitable**

Parametric models for covariance structure

Three sources of random variation in a typical set of longitudinal data:

- **Random effects**
- **Serial correlation**
- **Measurement error**
  - **when measurements involve delicate determinations, duplicate measurements at same time on same subject may show substantial variation**

Some simple models

- **Compound symmetry**

$$Y_{ij} - \mu_{ij} = U_i + Z_{ij}$$

$U_i \sim \mathrm{N}(0, \nu^2)$

$Z_{ij} \sim \mathrm{N}(0, \tau^2)$

Implies that $\mathrm{Corr}(Y_{ij}, Y_{ik}) = \nu^2/(\nu^2 + \tau^2)$, for all $j \neq k$

- **Random intercept and slope**

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

$(U_i, W_i) \sim \mathrm{BVN}(0, \Sigma)$

$Z_{ij} \sim \mathrm{N}(0, \tau^2)$

Often fits short sequences well, but extrapolation dubious, for example $\mathrm{Var}(Y_{ij})$ quadratic in t_{ij}

- **Autoregressive**

$$Y_{ij} - \mu_{ij} = \alpha(Y_{i,j-1} - \mu_{i,j-1}) + Z_{ij}$$

$Y_{i1} - \mu_{i1} \sim \mathrm{N}\{0, \tau^2/(1-\alpha^2)\}$

$Z_{ij} \sim \mathrm{N}(0, \tau^2), \quad j = 2, 3, \ldots$

Not a natural choice for underlying continuous-time processes

- **Stationary Gaussian process**

$$Y_{ij} - \mu_{ij} = W_i(t_{ij})$$

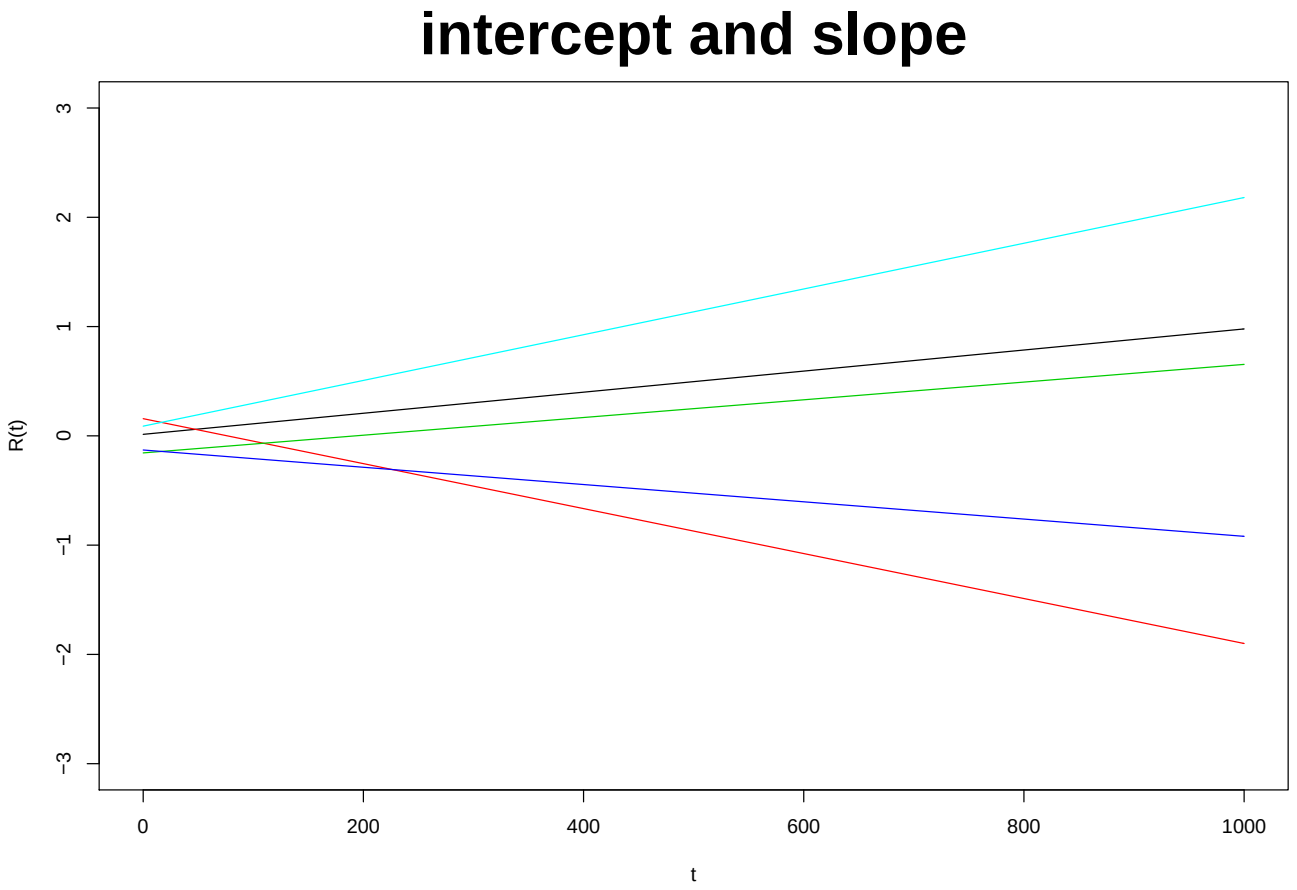
$W_i(t)$ a continuous-time Gaussian process

$\mathrm{E}[W(t)] = 0 \quad \mathrm{Var}\{W(t)\} = \sigma^2$

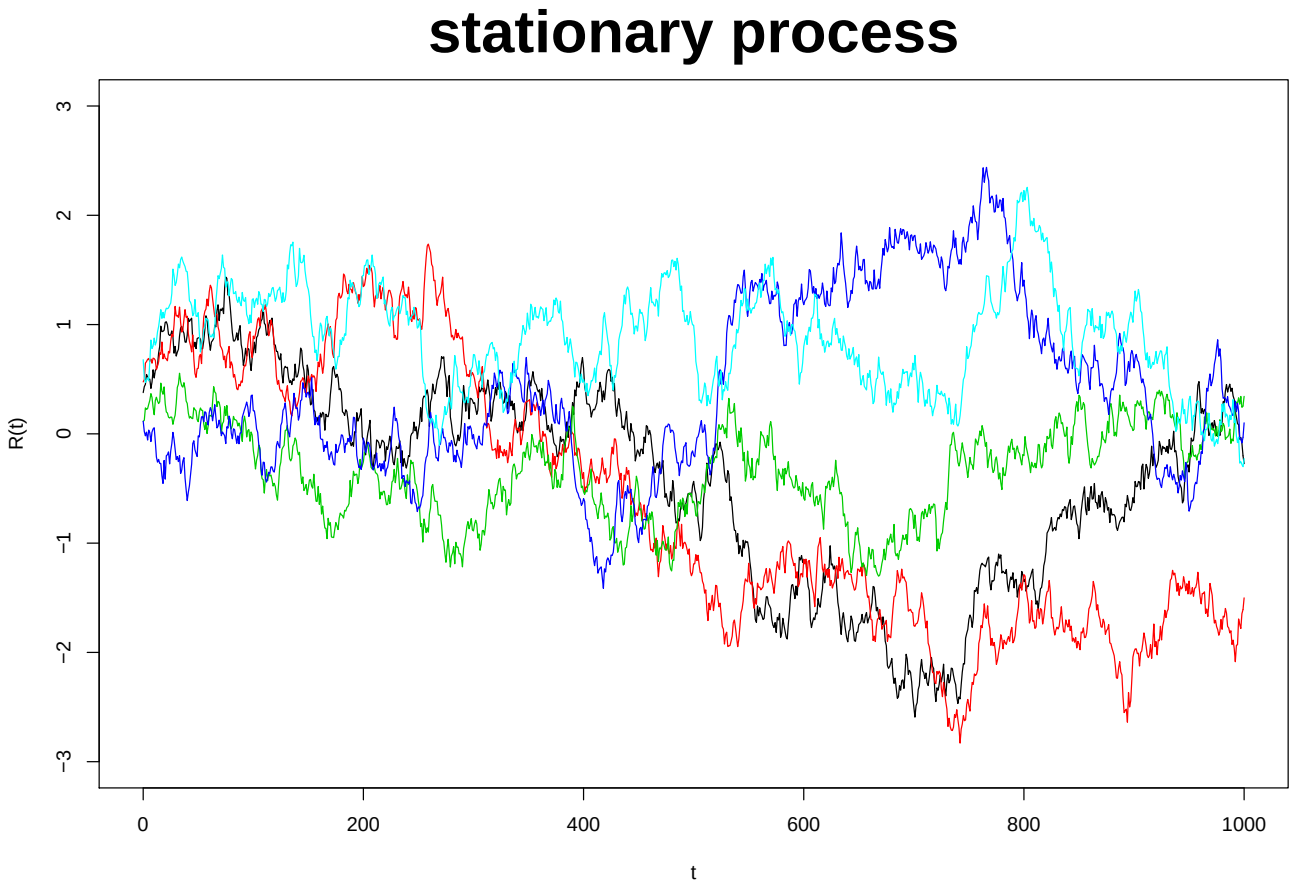
$\mathrm{Corr}\{W(t), W(t-u)\} = \rho(u)$

$\rho(u) = \exp(-u/\phi)$ gives continuous-time version of the autoregressive model

Time-varying random effects



Time-varying random effects: continued



- **A general model**

$$Y_{ij} - \mu_{ij} = d'_{ij} U_i + W_i(t_{ij}) + Z_{ij}$$

$U_i \sim \text{MVN}(0, \Sigma)$
(random effects)

d_{ij} = vector of explanatory variables for random effects

$W_i(t)$ = continuous-time Gaussian process
(serial correlation)

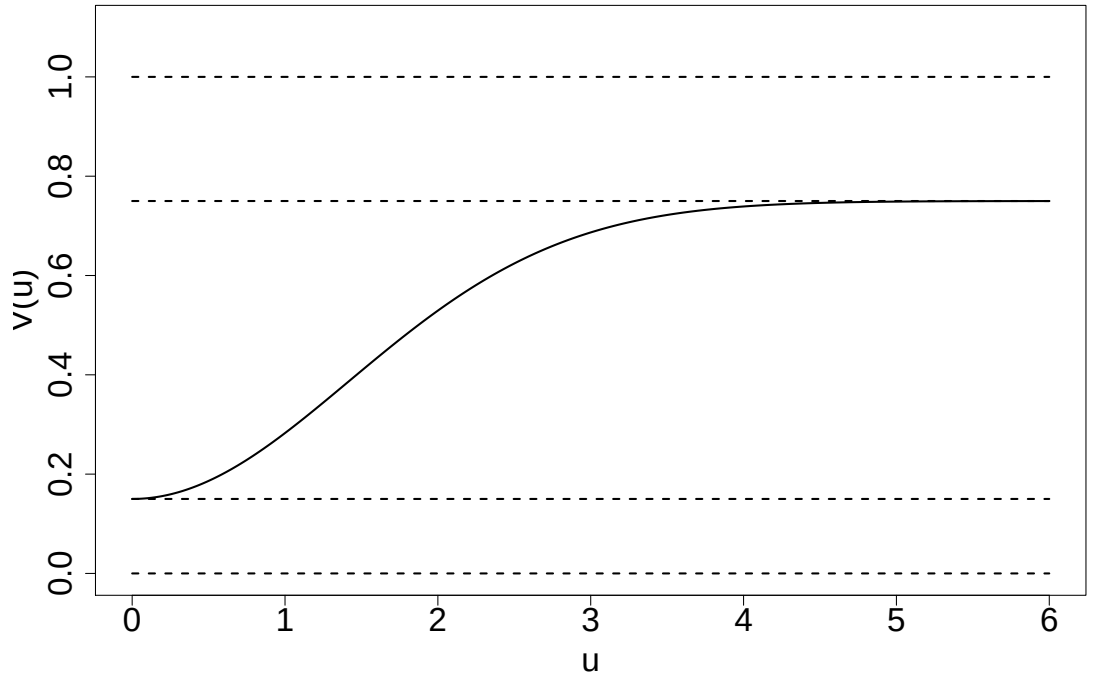
$Z_{ij} \sim \text{N}(0, \tau^2)$
(measurement errors)

Even when all three components of variation are needed in principle, one or two may dominate in practice

The variogram of the general model

$$Y_{ij} - \mu_{ij} = d'_{ij} U_i + W_i(t_{ij}) + Z_{ij}$$

$$V(u) = \tau^2 + \sigma^2\{1 - \rho(u)\} \quad \mathrm{Var}(Y_{ij}) = \nu^2 + \sigma^2 + \tau^2$$



Fitting the model

1. **A non-technical summary**
2. **The gory details**

Fitting the model: non-technical summary

- **Ad hoc methods won't do**
- **Likelihood-based inference is the statistical gold standard**
- **But be sure you know what you are estimating**

Fitting the model: the gory details

1. **The clever ostrich: robust version of ordinary least squares**
2. **The very clever ostrich: robust version of weighted least squares**
3. **Likelihood-based inference: ML and REML**

The clever ostrich

- use ordinary least squares for exploratory analysis and point estimation (ostrich)
- use sample covariance matrix of residuals to give consistent estimates of standard errors (clever)

Procedure as follows:

- $y = X\beta + \epsilon : Var(\epsilon) = V$
- $\hat{\beta} = (X'X)^{-1}X'y \equiv Dy$
- $\text{Var}(\hat{\beta}) = DVD' \simeq D\hat{V}D'$,
  $\hat{V}$ = sample covariance matrix of OLS residuals

$$\hat{\beta} \sim MVN(\beta, D\hat{V}D')$$

Good points:

- technically simple
- often reasonably efficient, and efficiency can be improved by using plausible weighting matrix W to reflect likely covariance structure (see below)
- don't need to specify covariance structure.

Bad points:

- sometimes very inefficient (recall linear regression example)
- accurate non-parametric estimation of V needs high replication (small n_i, large m)
- assumes missingness completely at random (more on this later)

Weighted least squares estimation

Weighted least squares estimate of β minimizes

$$S(\beta) = (y - X\beta)'W(y - X\beta)$$

where W is a symmetric weight matrix

Solution is

$$\tilde{\beta}_W = (X'WX)^{-1}X'Wy.$$

- **unbiased : $E(\tilde{\beta}_W) = \beta$, for any choice of W,**
- **$\mathrm{Var}(\tilde{\beta}_W) = \{(X'WX)^{-1}X'W\}V\{WX(X'WX)^{-1}\} = \Sigma$**

Inference

$$\tilde{\beta}_W \sim \mathrm{N}(\beta, \Sigma)$$

Special cases

1. $W = I$: ordinary least squares
   - $\tilde{\beta} = (X'X)^{-1}X'y$,
   - $\text{Var}(\tilde{\beta}) = (X'X)^{-1}X'VX(X'X)^{-1}$.
2. $W = V^{-1}$: maximum likelihood under Gaussian assumptions with known V
   - $\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y$,
   - $\text{Var}(\hat{\beta}) = (X'V^{-1}X)^{-1}$.

Maximum likelihood estimation (V_0 known)

Log-likelihood for observed data y is

$$\begin{aligned} L(\beta, \sigma^2, V_0) &= -0.5\{nm \log \sigma^2 + m \log |V_0| \\ &\quad + \sigma^{-2}(y - X\beta)'(I \otimes V_0)^{-1}(y - X\beta)\} \end{aligned} \tag{1}$$

where $I \otimes V_0$ is block-diagonal matrix, non-zero blocks V_0

Given V_0, estimator for β is

$$\hat{\beta}(V_0) = (X'(I \otimes V_0)^{-1}X)^{-1}X'(I \otimes V_0)^{-1}y, \tag{2}$$

the weighted least squares estimates with $W = (I \otimes V_0)^{-1}$.

Explicit estimator for σ^2 also available as

$$\hat{\sigma}^2(V_0) = RSS(V_0)/(nm) \tag{3}$$

$$RSS(V_0) = \{y - X\hat{\beta}(V_0)\}'(I \otimes V_0)^{-1}\{y - X\hat{\beta}(V_0)\}.$$

Maximum likelihood estimation, V_0 unknown

Substitute (2) and (3) into (1) to give reduced log-likelihood

$$\mathcal{L}(V_0) = -0.5m[n \log\{RSS(V_0)\} + \log|V_0|]. \qquad (4)$$

Numerical maximization of (4) then gives $\hat{V}_0$, hence $\hat{\beta} \equiv \hat{\beta}(\hat{V}_0)$ and $\hat{\sigma}^2 \equiv \hat{\sigma}^2(\hat{V}_0)$.

- **Dimensionality of optimisation is $\frac{1}{2}n(n+1) - 1$**
- **Each evaluation of $\mathcal{L}(V_0)$ requires inverse and determinant of an n by n matrix.**

REML: what is it and why use it?

- design matrix X influences estimation of covariance structure, hence
  - wrong X gives inconsistent estimates of σ^2 and V_0.
- remedy for designed experiments (n measurement times and g treatment groups) is to assume a saturated treatments-by-times model for estimation of σ^2 and V_0
- but
  - model for mean response then has ng parameters
  - if ng is large, maximum likelihood estimates of σ^2 and V_0 may be seriously biased
- saturated model not well-defined for most observational studies

Restricted maximum likelihood (REML)

- **REML is a generalisation of the unbiased sample variance estimator,**

$$s^2 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

- **Assume that Y follows a linear model as before,**

$$Y \sim MVN\{X\beta, \sigma^2(I \otimes V_0)\}.$$

- **transform data Y to $Y^* = Ay$, matrix A chosen to make distribution of Y^* independent of β.**

Example: transform to ordinary least squares residual space:

$$\tilde{\beta} = (X'X)^{-1}X'y \quad Y^* = Y - X\tilde{\beta} = \{I - X(XX)^{-1}X'\}Y$$

REML calculations

- $Y^* = Ay$
- Y^* has singular multivariate Normal distribution,

$$Y^* \sim MVN\{0, \sigma^2 A(I \otimes V_0)A'\}$$

  independent of β.

- estimate σ^2 and V_0 by maximising likelihood based on the transformed data Y^*.

$$\tilde{\beta}(V_0) = (X'(I \otimes V_0)^{-1}X)^{-1}X'(I \otimes V_0)^{-1}Y$$

$$\tilde{\sigma}^2(V_0) = RSS(V_0)/(N-p)$$

$$\begin{aligned}
\mathcal{L}^*(V_0) &= -0.5m[n\log\{RSS(V_0)\} + \log|V_0|] - 0.5\log|X'(I \otimes V_0)^{-1} \\
&= \mathcal{L}(V_0) - 0.5\log|X'(I \otimes V_0)^{-1}X|
\end{aligned}$$

Note that:

- different choices for A correspond to different rotations of coordinate axes within the residual space
- hence, REML estimates do not depend on A

```
fit1=lm(CD4~time)
summary(fit1)


library(nlme)
?lme
fit2=lme(CD4~time,random=~1|id)
summary(fit2)
```

Lecture 3

Generalized linear models for longitudinal data

- **marginal, transition and random effects models: why they address different scientific questions**
- **generalized estimating equations: what they can and cannot do**

Analysing non-Gaussian data

The classical GLM unifies previously disparate methodologies for a wide range of problems, including :

- **multiple regression/ANOVA (Gaussian responses)**
- **probit and logit regression (binary responses)**
- **log-linear modelling (categorical responses)**
- **Poisson regression (counted responses)**
- **survival analysis (non-negative continuous responses).**

How should we extend the classical GLM to analyse longitudinal data?

Generalized linear models for independent responses

Applicable to mutually independent responses $Y_i : i = 1, ..., n$.

1. **$E(Y_i) = \mu_i : h(\mu_i) = x_i'\beta$, where $h(\cdot)$ is known link function, x_i is vector of explanatory variables attached to i^{th} response, Y_i**

2. **$\text{Var}(Y_i) = \phi v(\mu_i)$ where $v(\cdot)$ is known variance function**

3. **pdf of Y_i is $f(y_i; \mu_i, \phi)$**

Two examples of classical GLM's

Example 1: simple linear regression

$$Y_i \sim N(\beta_1 + \beta_2 d_i, \sigma^2)$$

- $x_i = (1, d_i)'$
- $h(\mu) = \mu$
- $v(\mu) = 1$, $\phi = \sigma^2$
- $f(y_i; \mu_i, \phi) = \mathrm{N}(\mu_i, \phi)$

Example 2: Bernoulli logistic model (binary response)

$$P(Y_i = 1) = \exp(\beta_1 + \beta_2 d_i)/\{1 + \exp(\beta_1 + \beta_2 d_i)\}$$

- $x_i = (1, d_i)'$
- $h(\mu) = \log\{\mu/(1-\mu)$
- $v(\mu) = \mu(1-\mu)$, $\phi = 1$
- $f(y_i; \mu_i) =$ Bernoulli

Three GLM constructions for longitudinal data

- random effects models
- transition models
- marginal models

Random effects GLM

Responses $Y_{i1}, \ldots, Y_{in_i}$ on an individual subject conditionally independent, given unobserved vector of random effects U_i, $i = 1, \ldots, m$.

$U_i \sim g(\theta)$ represents properties of individual subjects that vary randomly between subjects

- $E(Y_{ij}|U_i) = \mu_{ij} : h(\mu_{ij}) = \mathbf{x}'_{ij}\beta + \mathbf{z}'_{ij}U_i$
- $\text{Var}(Y_{ij}|U_i) = \phi v(\mu_{ij})$
- $(Y_{i1}, \ldots, Y_{in_i})$ are mutually independent conditional on U_i.

Likelihood inference requires evaluation of

$$f(y) = \int \prod_{i=1}^{m} f(y_i|U_i)g(U_i)dU_i$$

Transition GLM

Conditional distribution of each Y_{ij} modelled directly in terms of preceding $Y_{i1}, \ldots, Y_{ij-1}$.

- $E(Y_{ij}|\text{history}) = \mu_{ij}$
- $h(\mu_{ij}) = \mathbf{x}'_{ij}\beta + Y'_{(ij)}\alpha$, where $Y_{(ij)} = (Y_{i1}, \ldots, Y_{ij-1})$
- $\text{Var}(Y_{ij}|\text{history}) = \phi v(\mu_{ij})$

Construct likelihood as product of conditional distributions, usually assuming restricted form of dependence, for example:

$$f_k(y_{ij}|y_{i1}, ..., y_{ij-1}) = f_k(y_{ij}|y_{ij-1})$$

and condition on y_{i1} as model does not directly specify $f_{i1}(y_{i1})$.

Marginal GLM

Let $h(\cdot)$ be a link function which operates component-wise,

- $E(y_{ij}) = \mu_{ij} : h(\mu_{ij}) = X_{ij}'\beta$
- $\text{Var}(y_{ij}) = \phi v(\mu_{ij})$
- $\text{Corr}(y_{ij}) = R(\alpha)_{ij}$.

Not a fully specified probability model

May require constraints on variance function $v(\cdot)$ and correlation matrix $R(\cdot)$ for valid specification

Inference for β uses method of generalized estimating equations (the clever ostrich revisited)

Indonesian children's health study

- **ICHS - of interest is to investigate the association between risk of respiratory illness and vitamin A deficiency**
- **Over 3000 children medically examined quarterly for up to six visits to assess whether they suffered from respiratory illness (yes/no = 1/0) and xerophthalmia (an ocular manifestation of vitamin A deficiency) (yes/no = 1/0).**
- **Let Y_{ij} be the binary random variable indicating whether child i suffers from respiratory illness at time t_{ij}.**
- **Let x_{ij} be the covariate indicating whether child i is vitamin A deficient at time t_{ij}.**

Marginal model for ICHS study

- $\mathbf{E}[Y_{ij}] = \mu_{ij} = \mathbf{P}P(Y_{ij} = 1)$
- $\log(\frac{\mu_{ij}}{1-\mu_{ij}}) = \beta_0 + \beta_1 x_{ij}$
- $\mathbf{Var}(Y_{ij}) = \mu_{ij}(1 - \mu_{ij})$
- $\mathbf{Corr}(Y_{ij}, Y_{ik}) = \alpha.$

Marginal model - interpretation of regression parameters

- $\exp(\beta_0)$ is the odds of infection for any child with replete vitamin A.
- $\exp(\beta_1)$ is the odds ratio for any child - i.e. the odds of infection among vitamin A deficient children divided by the odds of infection among children replete with vitamin A.
- $\exp(\beta_1)$ is a ratio of population frequencies - a population-averaged parameter.
- β_1 represents the effect of the explanatory variable (vitamin A status) on any child's chances of respiratory infection.

Random effects model for ICHS study

- $\mathbf{E}[Y_{ij}|U_i] = \mu_{ij} = \mathbf{P}(Y_{ij} = 1|U_i)$
- $\log(\frac{\mu_{ij}}{1-\mu_{ij}}) = \beta_0^* + U_i + \beta_1^* x_{ij}$ where $U_i \sim N(0, \gamma^2)$
- U_i represents the i^{th} child's propensity for infection attributed to unmeasured factors (which could be genetic,
  environmental ...)
- $\mathbf{Var}(Y_{ij}|U_i) = \mu_{ij}(1 - \mu_{ij})$
- $Y_{ij}|U_i \perp Y_{ik}|U_i$ for $j \neq k$.

Random effects model - interpretation of regression parameters

- $\exp(\beta_0^*)$ is the odds of infection for a child with average propensity for infection and with replete vitamin A.
- $\exp(\beta_1^*)$ is the odds ratio for a specific child - i.e. the odds of infection for a vitamin A deficient child divided by the odds of infection for the same child replete with vitamin A.
- β_1 represents the effect of the explanatory variable (vitamin A status) upon an individual child's chance of respiratory infection.

Estimating equations

Estimating equations for β in a classical GLM:

$$S(\beta_j) = \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \beta_j} v_i^{-1} (Y_i - \mu_i) = 0 : j = 1, ..., p$$

where $v_i = \text{Var}(Y_i)$.

In vector-matrix notation:

$$S(\beta) = D'_{\mu\beta} V^{-1} (Y - \mu) = 0$$

- **$D_{\mu\beta}$ is an $n \times p$ matrix with ij^{th} element $\frac{\partial \mu_i}{\partial \beta_j}$**
- **V is an $n \times n$ diagonal matrix with non-zero elements proportional to $\text{Var}(Y_i)$**
- **Y and μ are n-element vectors with elements Y_i and μ_i**

Generalized estimating equations (GEE)

In longitudinal setting:

- **in previous slide Y_i and μ_i were scalars. In the longitudinal setting they are replaced by n_i-element vectors Y_i and μ_i, associated with i^{th} subject**
- **corresponding matrices $V_i(\alpha) = \text{Var}(Y_i)$ are no longer diagonal**

Estimating equations for complete set of data, $Y = (Y_1, ..., Y_m)$,

$$S(\beta) = \sum_{i=1}^{m} \{D_{\mu_i \beta}\}' \{V_i(\alpha)\}^{-1} (Y_i - \mu_i) = 0$$

Large-sample properties of resulting estimates $\hat{\beta}$

$$\sqrt{(m)}(\hat{\beta} - \beta) \sim MVN(0, I_0^{-1}) \tag{5}$$

where

$$I_0 = \sum_{i=1}^{m} \{D_{\mu_i \beta}\}' \{V_i(\alpha)\}^{-1} D_{\mu_i \beta}$$

What to do when variance matrices $V_i(\alpha)$ are unknown?

The working covariance matrix

$$S(\beta) = \sum_{i=1}^{m} \{D_{\mu_i \beta}\}' \{V_i^*(\alpha)\}^{-1} (Y_i - \mu_i) = 0$$

$V_i^*(\cdot)$ is a guess at the covariance matrix of Y_i, called the working covariance matrix

Result (5) on distribution of $\hat{\beta}$ now modified to

$$\sqrt{(m)}(\hat{\beta} - \beta) \sim MVN(0, I_0^{-1} I_1 I_0^{-1}) \tag{6}$$

where

$$I_0 = \sum_{i=1}^{m} \{D_{\mu_i \beta}\}' \{V_i(\alpha)\}^{-1} D_{\mu_i \beta}$$

and

$$I_1 = \sum_{i=1}^{m} \{D_{\mu_i \beta}\}' \{V_i^*(\alpha)\}^{-1} \mathrm{Var}(Y_i) \{V_i^*(\alpha)\}^{-1} D_{\mu_i \beta}$$

Properties:

- result (6) reduces to (5) if $V_i^*(\cdot) = V_i(\cdot)$
- estimator $\hat{\beta}$ is consistent even if $V_i^*(\cdot) \neq V_i(\cdot)$
- to calculate an approximation to I_1, replace $\text{Var}(Y_i)$ by

$$(Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)'$$

  where $\hat{\mu}_i = \mu_i(\hat{\beta})$

  Gives a terrible estimator of $\text{Var}(Y_i)$, but OK in practice provided:

  - number of subjects, m, is large
  - same model for μ_i fitted to groups of subjects;
  - observation times common to all subjects

- but a bad choice of $V_i^*(\cdot)$ does affect efficiency of $\hat{\beta}$.

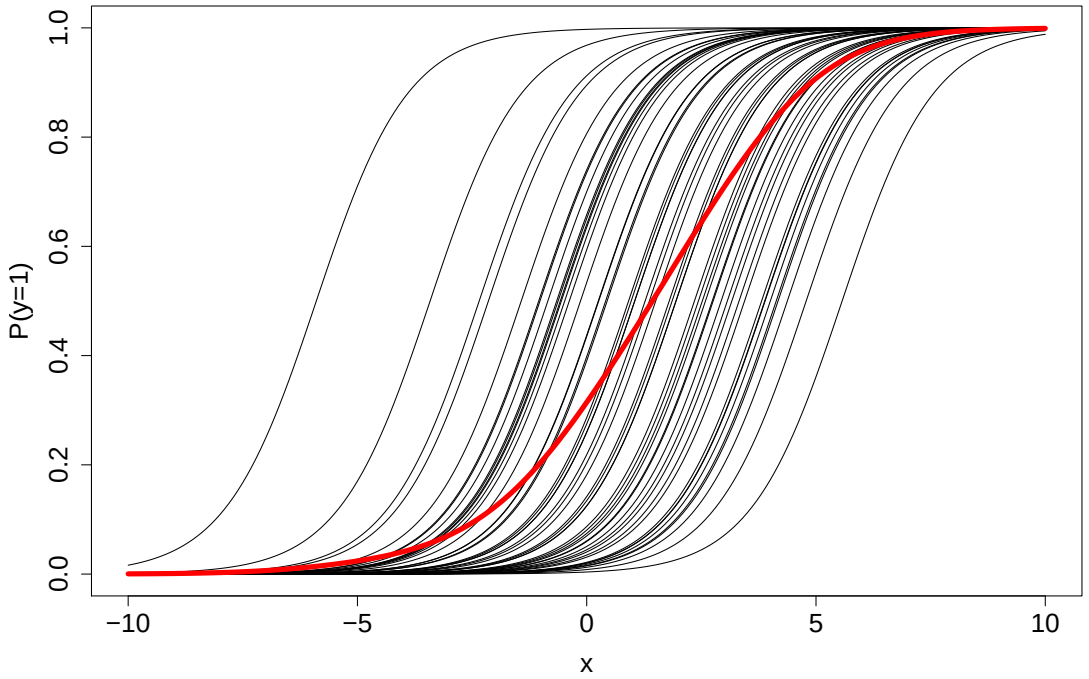
What are we estimating?

- in marginal modelling, β measures population-averaged effects of explanatory variables on mean response
- in transition or random effects modelling, β measures effects of explanatory variables on mean response of an individual subject, conditional on
  - subject's measurement history (transition model)
  - subject's own random characteristics U_i (random effects model)

Example: **Simulation of a logistic regression model, probability of positive response from subject i at time t is $p_i(t)$,**

$$\text{logit}\{p_i(t)\} : \alpha + \beta x(t) + \gamma U_i,$$

$x(t)$ is a continuous covariate and U_i is a random effect



Example: Effect of mother's smoking on probability of intra-uterine growth retardation (IUGR).

Consider a binary response $Y = 1/0$ to indicate whether a baby experiences IUGR, and a covariate x to measure the mother's amount of smoking.

Two relevant questions:

1. **public health: by how much might population incidence of IUGR be reduced by a reduction in smoking?**

2. **clinical/biomedical: by how much is a baby's risk of IUGR reduced by a reduction in their mother's smoking?**

Question 1 is addressed by a marginal model, question 2 by a random effects model

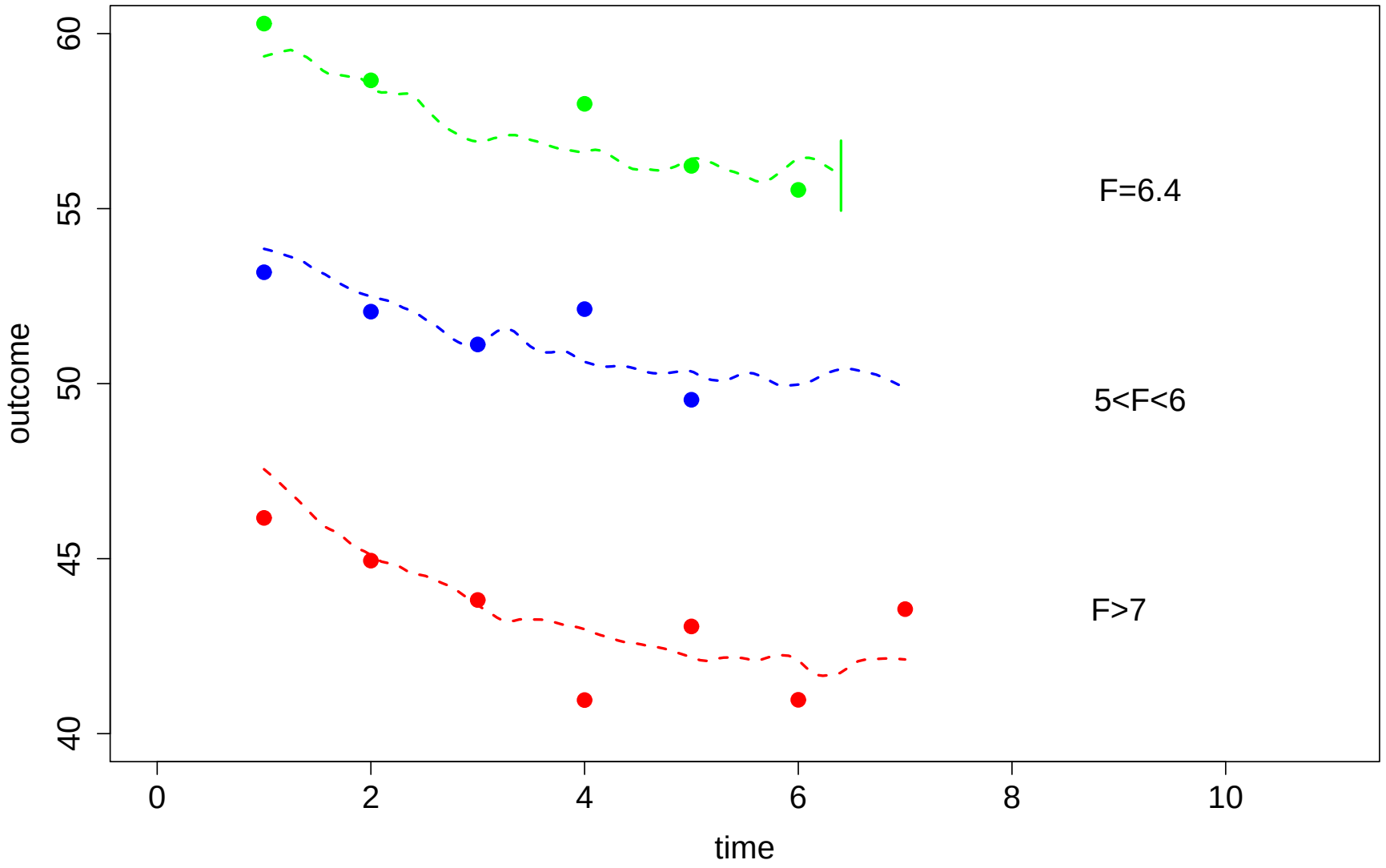
```
set.seed(2346)
x=rep(1:10,50)
logit=0.1*(x-mean(x))
subject=rep(1:50,each=10)
re=2*rnorm(50)
re=rep(re,each=10)
prob=exp(re+logit)/(1+exp(re+logit))
y=(runif(500)<prob)
fit1=glm(y~x,family=binomial)
summary(fit1)

library(gee)
fit2<-gee(y~x,id=subject,family=binomial)
summary(fit2)

library(glmmML)
fit3<-glmmML(y~x,family=binomial,cluster=subject)
summary(fit3)
```

Lecture 4.

- Dropouts
  - classification of missing value mechanisms
  - modelling the missing value process
  - what are we estimating?
- Joint modelling
  - what is it?
  - why do it?
  - random effects models
  - transformation models


outcome
60
55
50
45
40
0
2
4
6
8
10
time
F=6.4
5<F<6
F>7

Missing values and dropouts

Issues concerning missing values in longitudinal data can be addressed at two different levels:

- **technical: can the statistical method I am using cope with missing values?**
- **conceptual: *why* are the data missing? Does the fact that an observation is missing convey partial information about the value that would have been observed?**

These same questions also arise with cross-sectional data, but the correlation inherent to longitudinal data can sometimes be exploited to good effect.

Rubin's classification

- **MCAR (completely at random): P(missing) depends neither on observed nor unobserved measurements**
- **MAR (at random): P(missing) depends on observed measurements, but not on unobserved measurements conditional on observed measurements**
- **MNAR (not at random): conditional on observed measurements, P(missing) depends on unobserved measurements.**

Example : Longitudinal clinical trial

- **completely at random: patient leaves the the study because they move house**
- **at random : patient leaves the study on their doctor's advice, based on observed measurement history**
- **not at random : patient misses their appointment because they are feeling unwell.**

Intermittent missing values and dropouts

- **dropouts: subjects leave study prematurely, and never come back**
- **intermittent missing values: everything else**

Sometimes reasonable to assume intermittent missing values are also missing completely at random

Not so for dropouts

It is always helpful to know *why* subjects drop out

Modelling the missing value process

- $Y = (Y_1, ..., Y_n)$, intended measurements on a single subject
- $t = (t_1, ..., t_n)$, intended measurement times
- $M = (M_1, ..., M_n)$, missingness indicators
- for dropout, M reduces to a single dropout time D, in which case:
  - $(Y_1, ..., Y_{D-1})$ observed
  - $(Y_D, ..., Y_n)$ missing

A model for data subject to missingness is just a specification of the joint distribution

$$[Y, M]$$

Modelling the missing value process: three approaches

- **Selection factorisation**

$$[\boldsymbol{Y}, \boldsymbol{M}] = [\boldsymbol{Y}][\boldsymbol{M}|\boldsymbol{Y}]$$

- **Pattern mixture factorisation**

$$[\boldsymbol{Y}, \boldsymbol{M}] = [\boldsymbol{M}][\boldsymbol{Y}|\boldsymbol{M}]$$

- **Random effects**

$$[\boldsymbol{Y}, \boldsymbol{M}] = \int [\boldsymbol{Y}|\boldsymbol{U}][\boldsymbol{M}|\boldsymbol{U}][\boldsymbol{U}]\boldsymbol{dU}$$

Comparing the three approaches

- **Pattern mixture factorisation has a natural data-analytic interpretation**
  (sub-divide data into different dropout-cohorts)
- **Selection factorisation may have a more natural mechanistic interpretation in the dropout setting**
  (avoids conditioning on the future)
- **Random effects conceptually appealing, especially for noisy measurements, but make stronger assumptions and usually need computationally intensive methods for likelihood inference**

Fitting a model to data with dropouts

- **MCAR**

  1. **almost any method will give sensible point estimates of mean response profiles**
  2. **almost any method which takes account of correlation amongst repeated measurements will give sensible point estimates and standard errors**

- **MAR**

    1. **likelihood-based inference implicitly assumes MAR**
    2. **for inferences about a hypothetical dropout-free population, there is no need to model the dropout process explicitly**
    3. **but be sure that a hypothetical dropout-free population is the required target for inference**

- **MNAR**

  1. **joint modelling of repeated measurements and dropout times is (more or less) essential**
  2. **but inferences are likely to be sensitive to modelling assumptiuons that are difficult (or impossible) to verify empirically**

Longitudinal data with dropouts: the gory details

New notation for measurements on a single subject:

- $Y^* = (Y_1^*, \ldots, Y_n^*)$: complete intended sequence
- $t = (t_1, \ldots, t_n)$: times of intended measurements
- $Y = (Y_1, \ldots, Y_n)$: incomplete observed sequence
- $H_k = \{Y_1, \ldots, Y_{k-1}\}$: observed history up to time t_{k-1}

Core assumption:

$$Y_k = \begin{cases} Y_k^* & : \quad k = 1, 2, \ldots, D-1 \\ 0 & : \quad k \geq D \end{cases}$$

No *a priori* separation into sub-populations of potential dropouts and non-dropouts

The likelihood function

Two basic ingredients of any model:

1. $y^* \sim f^*(y; \beta, \alpha)$,
2. $P(D = d|\text{history}) = p_d(H_d, y_d^*; \phi)$.

- β parameterises mean response profile for y^*
- α parameterises covariance structure of y^*
- ϕ parameterises dropout process.

For inference, need the likelihood for the *observed* data, y, rather than for the *intended* data y^*

Let $f_k^*(y|H_k;\beta,\alpha)$ denote conditional pdf of Y_k^* given H_k

Model specifies $f_k^*(\cdot)$, we need $f_k(\cdot)$.

1. $P(Y_k = 0|H_k, Y_{k-1} = 0) = 1$

   because dropouts never re-enter the study.

2. $$\mathrm{P}(Y_k = 0|H_{k-1}, Y_{k-1} \neq 0) = \int p_k(H_k, y;\phi) f_k^*(y|H_k;\beta,\alpha)dy$$

3. For $Y_k \neq 0$,

   $$f_k(y|H_k;\beta,\alpha,\phi) = \{1 - p_k(H_k, y;\phi)\} f_k^*(y|H_k;\beta,\alpha).$$

Multiply sequence of conditional distributions for Y_k given H_k to define joint distribution of Y, and hence likelihood function

1. **for a complete sequence $Y = (Y_1, \ldots, Y_n)$:**

$$f(y) = f^*(y) \prod_{k=2}^{n} \{1 - p_k(H_k, y_k)\}$$

2. **for an incomplete sequence $Y = (Y_1, \ldots, Y_{d-1}, 0, \ldots, 0)$:**

$$f(y) = f^*_{d-1}(y) \prod_{k=2}^{d-1} \{1 - p_k(H_k, y_k)\} \mathrm{P}(Y_d = 0 | H_d, Y_{d-1} \neq 0)$$

where $f^*_{d-1}(y)$ denotes joint pdf of $(Y_1^*, ..., Y_{d-1}^*)$.

Now consider a set of data with m subjects.

- β and α parameterise measurement process y^*
- ϕ parameterises dropout process

Hence, log-likelihood can be partitioned into three components:

$$L(\beta, \alpha, \phi) = L_1(\beta, \alpha) + L_2(\phi) + L_3(\beta, \alpha, \phi)$$

$$L_1(\beta, \alpha) = \sum_{i=1}^{m} \log\{f^*_{d_i - 1}(y_i)\} \quad L_2(\phi) = \sum_{i=1}^{m} \sum_{k=1}^{d_i - 1} \log\{1 - p_k(H_{ik}, y_{ik})\}$$

$$L_3(\beta, \alpha, \phi) = \sum_{i : d_i \leq n} \log\{\mathrm{P}(Y_{id_i} = 0 | H_{id_i} Y_{id_i - 1} \neq 0)\}.$$

When is likelihood inference straightforward?

$$L_3(\beta, \alpha, \phi) = \sum_{i:d_i \leq n} \log\{\mathrm{P}(Y_{id_i} = 0 | H_{id_i} Y_{id_i - 1} \neq 0).$$

If $L_3(\cdot)$ only depends on ϕ, inference is straightforward, because we can then:

- absorb $L_3(\cdot)$ into $L_2(\cdot)$
- maximise $L_1(\beta, \alpha)$ and $L_2(\phi)$ separately

$$L_3(\beta, \alpha, \phi) = \sum_{i:d_i \leq n} \log\{\mathrm{P}(Y_{id_i} = 0 | H_{id_i} Y_{id_{i-1}} \neq 0).$$

- $\mathrm{P}(Y_k = 0 | H_{k-1}, Y_{k-1} \neq 0) = \int p_k(H_k, y; \phi) f_k^*(y | H_k; \beta, \alpha) dy$
- MAR implies $p_k(H_k, y; \phi) = p_k(H_k; \phi)$ does not depend on y
- It follows that

$$\begin{aligned} \mathrm{P}(Y_k = 0 | H_{k-1}, Y_{k-1} \neq 0) &= p_k(H_k; \phi) \int f_k^*(y | H_k; \beta, \alpha) dy \\ &= p_k(H_k; \phi), \end{aligned}$$

since conditional pdf must integrate to one.

Key result:

- **If dropouts are MAR, then $L_3(\beta, \alpha, \phi) = L_3(\phi)$ and parameter estimates for the model can be obtained by separate maximisation of:**
  - $L_1(\beta, \alpha)$
  - $L_2^*(\phi) \equiv L_2(\phi) + L_3(\phi)$

Is MAR ignorable?

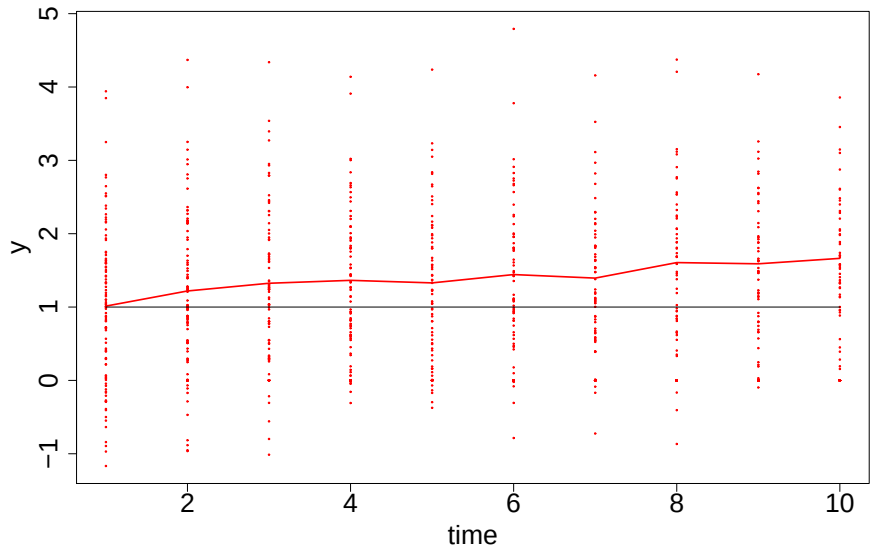
Conventional wisdom: if dropout is MAR and we only want estimates of β and α we can ignore the dropout process

Two caveats:

- If MAR holds, but measurement and dropout models have parameters in common, ignoring dropouts is potentially inefficient
- More importantly, parameters of the measurement model may not be the most appropriate target for inference

Example: simulated MAR data

- **Y*-process:** mean response $\mu(t) = 1$, constant correlation ρ between any two measurements on same subject.
- **dropout sub-model:** $\text{logit}(p_{ij}) = \alpha + \beta y_{ij-1}$
- **simulated realisation** for $\rho = 0.9$, $\alpha = -1$ and $\beta = -2$



In the simulation:

- **empirical means show a steadily rising trend**
- **likelihood analysis ignoring dropout concludes that mean response is constant over time.**

Explanation:

- **empirical means are estimating conditional expectation,**

$$E(Y^*(t)|\text{dropout time} > t)$$

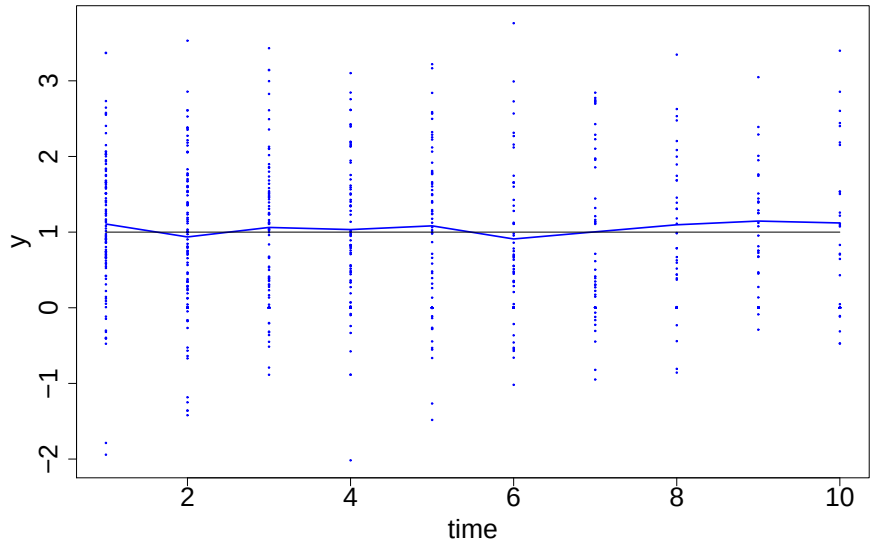
- **likelihood analysis is estimating unconditional expectation**

$$E[Y^*(t)]$$

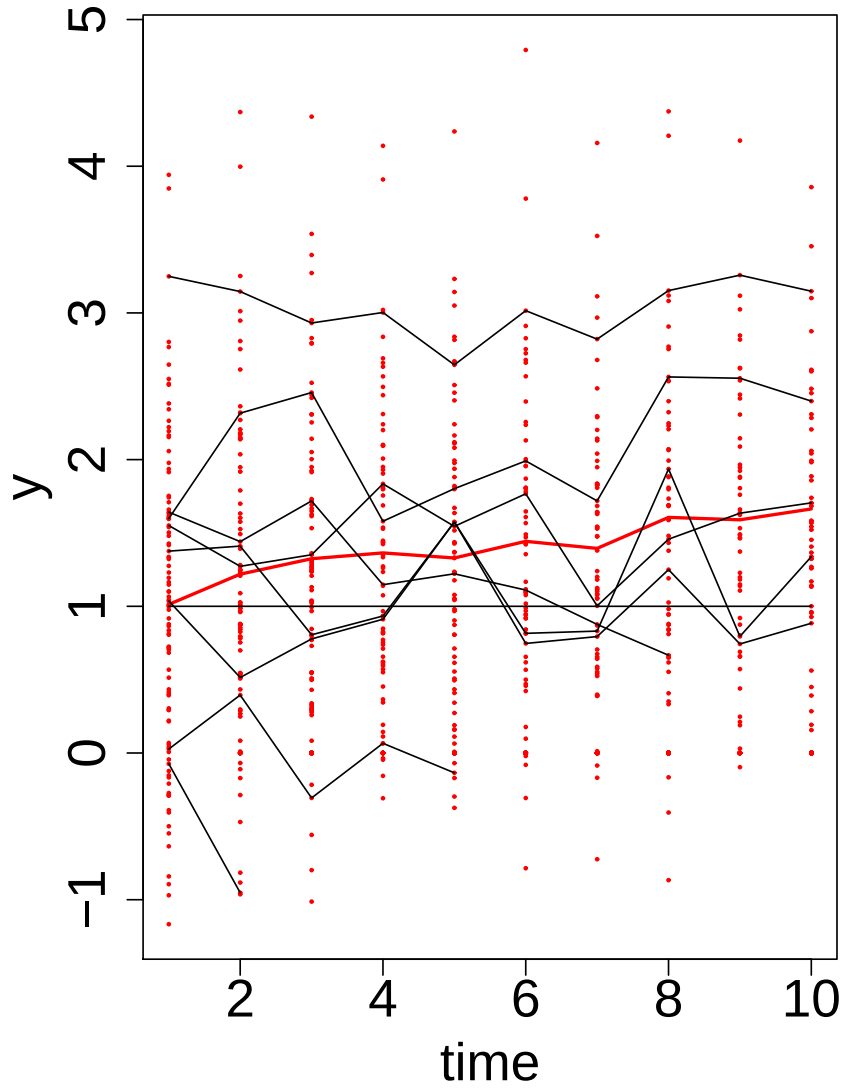
Which, if either, of these do you want to estimate?

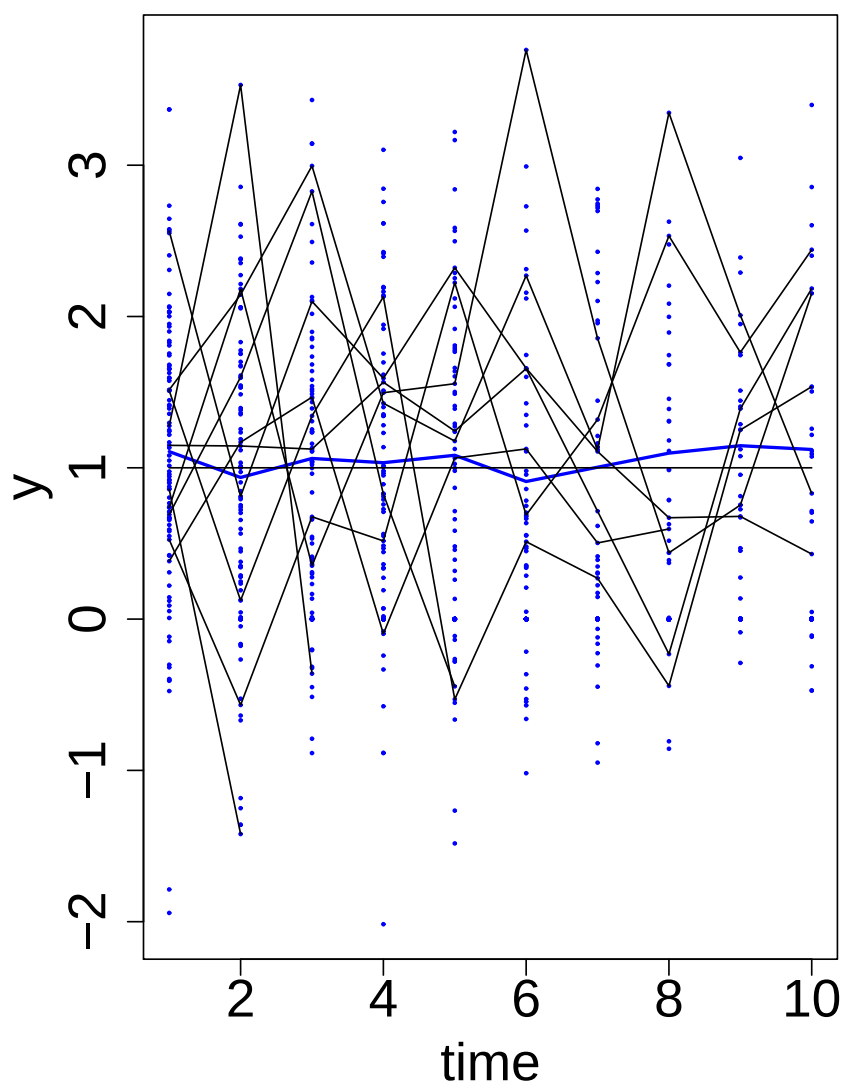
Under random dropout, conditional and unconditional means are different because the data are correlated.

Diagram below shows simulation with $\rho = 0$, i.e. no correlation, but $\alpha = -1$ and $\beta = -2$ as before.



Empirical means now tell same story as likelihood analysis, namely that mean response is constant over time.


y
5
4
3
2
1
0
−1
2
4
6
8
10
time


y
3
2
1
0
−1
−2
2
4
6
8
10
time

PJD's take on ignorability

For correlated data, dropout mechanism can be ignored only if dropouts are completely random

In all other cases, need to:

- **think carefully what are the relevant practical questions,**
- **fit an appropriate model for both measurement process and dropout process**
- **use the model to answer the relevant questions.**

Joint modelling: what is it?

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

$$[\boldsymbol{Y}, \boldsymbol{F} | \boldsymbol{x}, \boldsymbol{\theta}]$$

Prothrombin index data

- **Placebo-controlled RCT of prednisone for liver cirrhosis patients.**

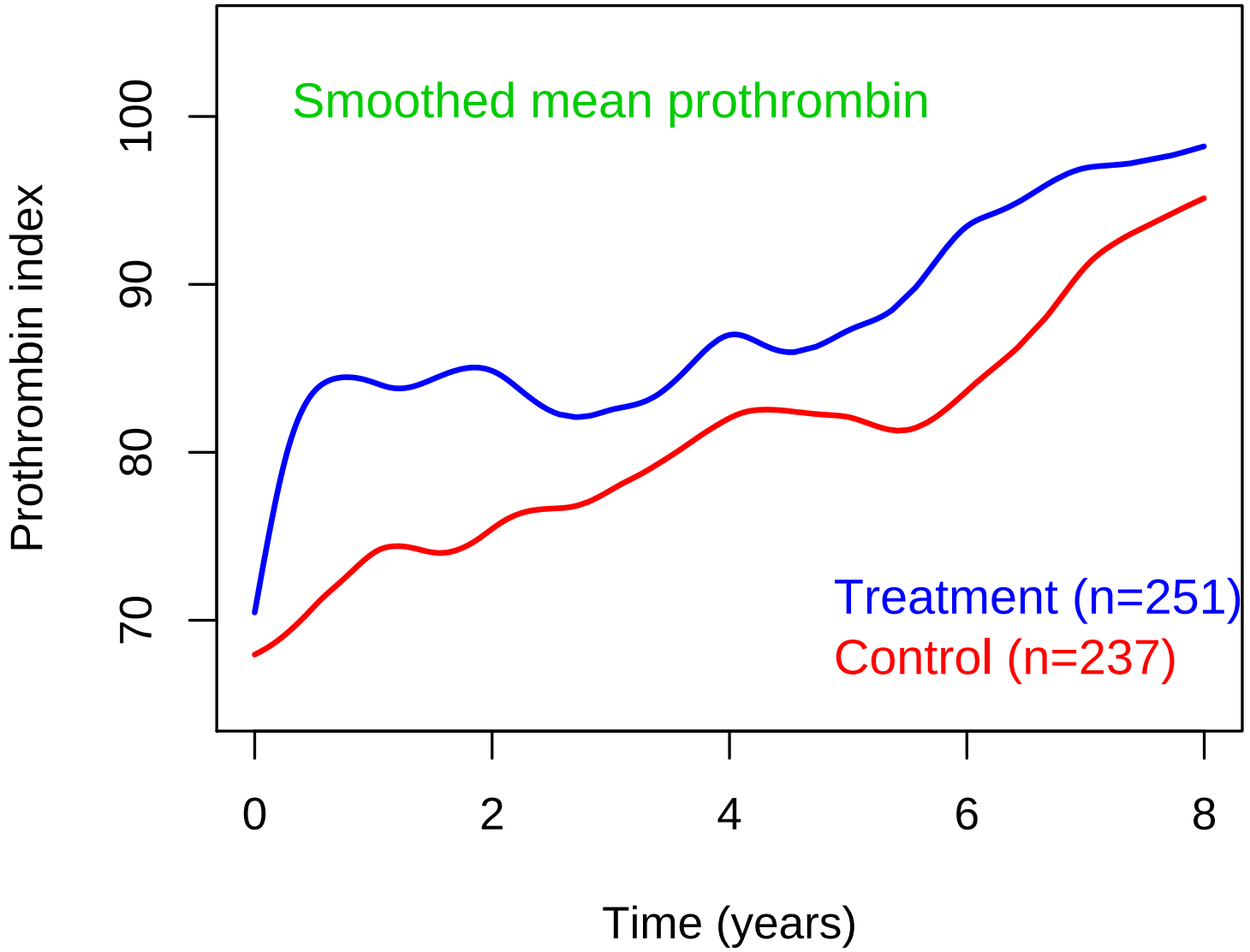
  Total $m = 488$ subjects.

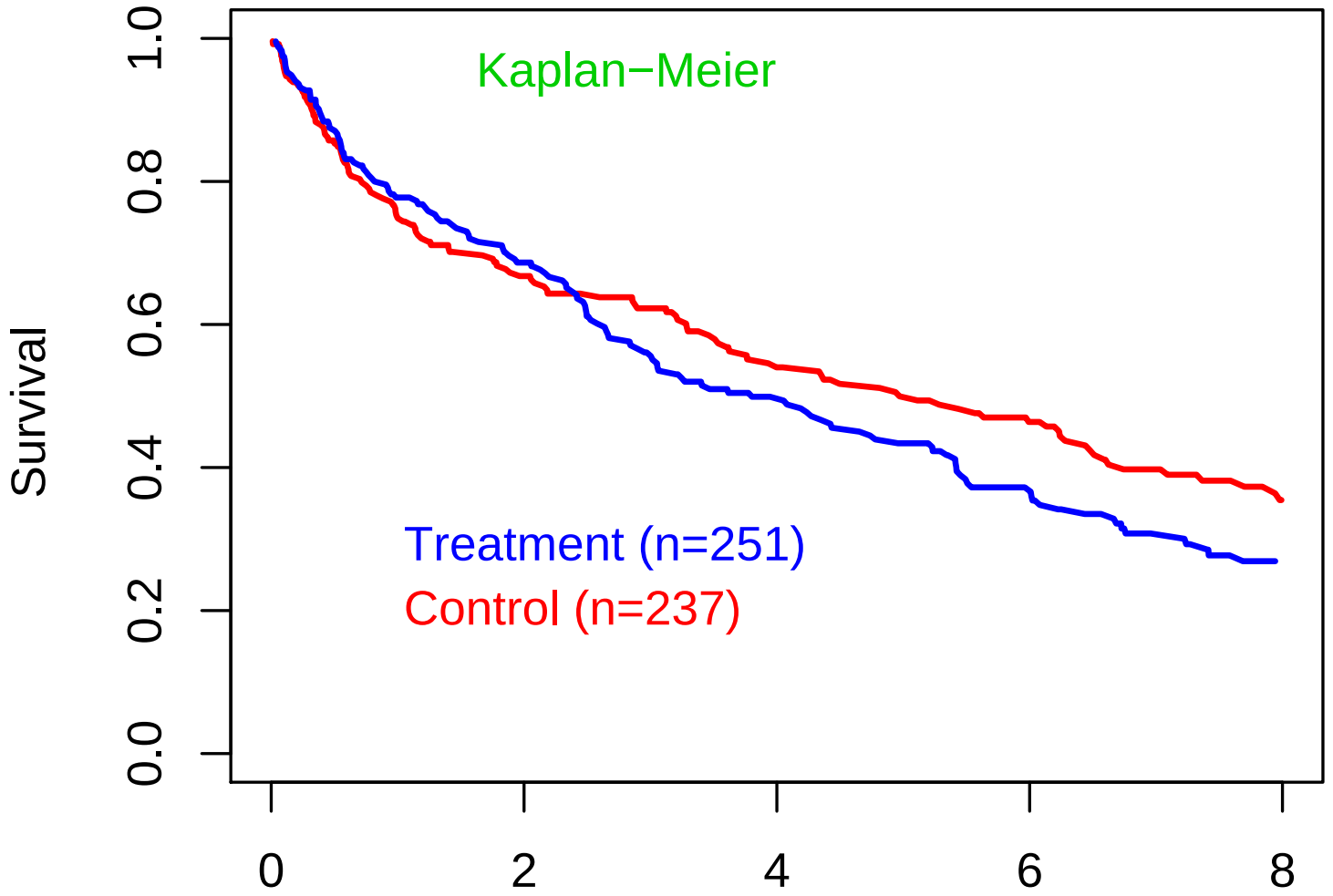
- **F = time of death**

  Y = time-sequence of prothrombin index measurements (months $\approx$ 0, 3, 6, 12, 24, 36,...,96)

- **$\approx$ 30% survival to 96 months**

Andersen, Borgan, Gill and Keiding, 1993


Smoothed mean prothrombin
Treatment (n=251)
Control (n=237)
Prothrombin index
70
80
90
100
0
2
4
6
8
Time (years)


Kaplan−Meier
Treatment (n=251)
Control (n=237)
Survival
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8

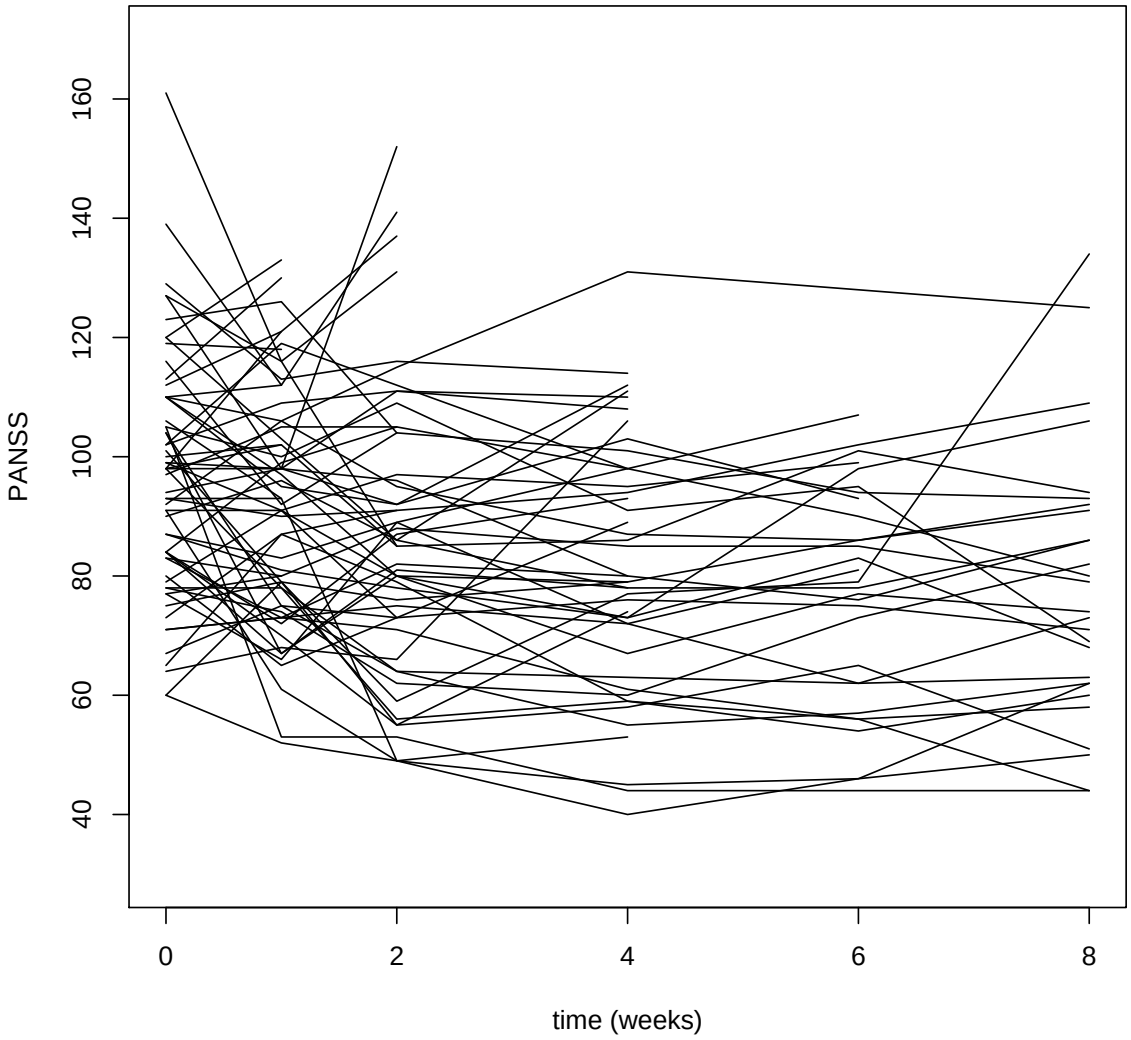
Schizophrenia trial data

- Data from placebo-controlled RCT of drug treatments for schizophrenia:
  - Placebo; Haloperidol (standard); Risperidone (novel)
- Y = sequence of weekly PANSS measurements
- F = dropout time
- Total $m = 516$ subjects, but 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 ($P > H > R$)

Schizophrenia data
PANSS responses from haloperidol arm

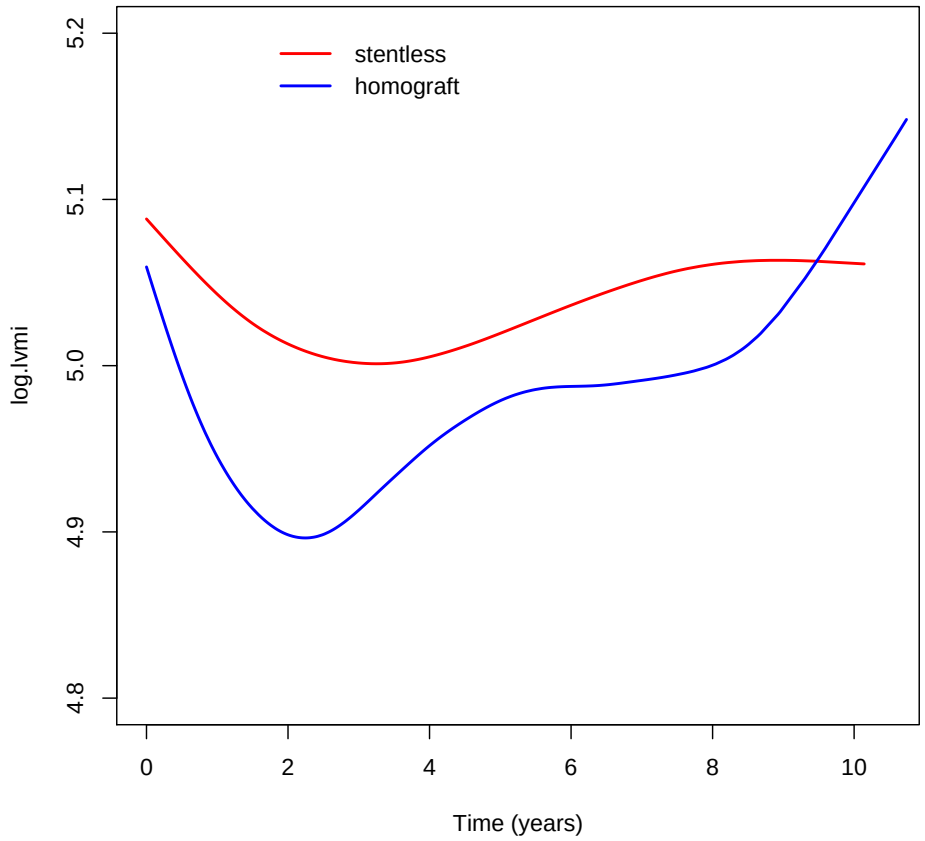


Heart surgery data

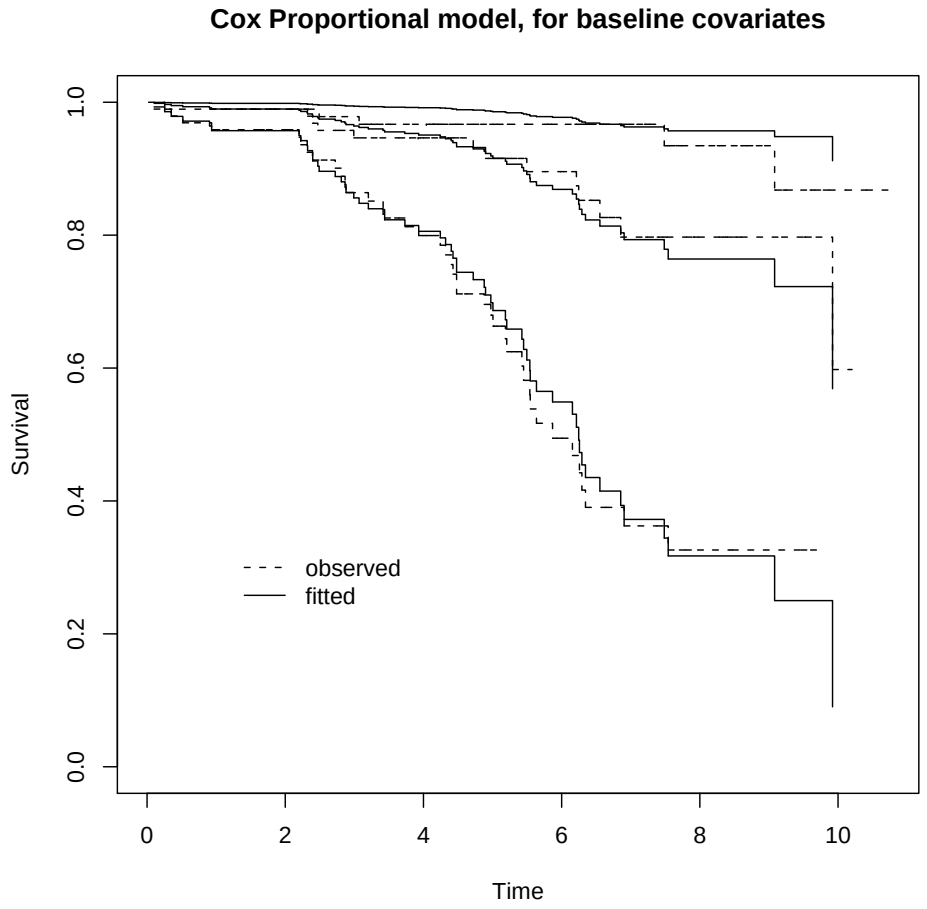
- Data from RCT to compare efficacy of two types of artificial heart-valves
  - homograft; stentless
- $m = 289$ subjects
- $Y =$ time-sequence of left-ventricular-mass-index (LVMI)
- $F =$ time of death
- two other repeated measures of heart-function also available (ejection fraction, gradient)

Lim et al, 2008

Heart surgery data
Mean log-LVMI response profiles



Heart surgery data
Survival curves adjusted for baseline covariates



Joint modelling: why do it?

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

Example: prothrombin index data

- **interest is in time to progression/death**
- **but slow progression of disease implies heavy censoring**
- **hence, joint modelling improves inferences about marginal distribution $[F]$**

Joint modelling: why do it?

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

Example: Schizophrenia data

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

Joint modelling: why do it?

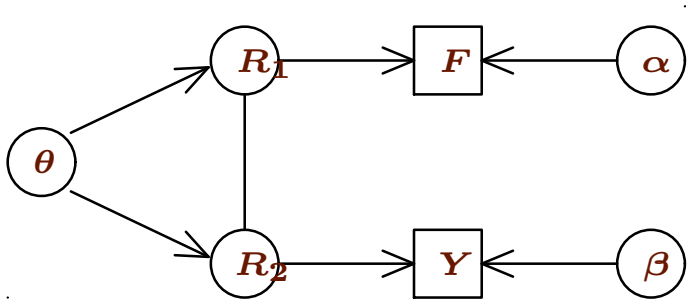
Because relationship between Y and F is of direct interest

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**
- **also interested in inter-relationships amongst LVMI, ejection fraction, gradient and survival time**

Random effects models

- linear Gaussian sub-model for repeated measurements
- proportional hazards sub-model with time-dependent fraility for time-to-event
- sub-models linked through shared random effects



Example: Henderson, Diggle and Dobson, 2000

Ingredients of model are:

- **a latent stochastic process; a measurement sub-model; a hazard sub-model**

Latent stochastic process

Bivariate Gaussian process $R(t) = \{R_1(t), R_2(t)\}$

- $R_k(t) = D_k(t)U_k + W_k(t)$
- $\{W_1(t), W_2(t)\}$: **bivariate stationary Gaussian process**
- (U_1, U_2): **multivariate Gaussian random effects**

Bivariate process $R(t)$ realised independently between subjects

Measurement sub-model

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

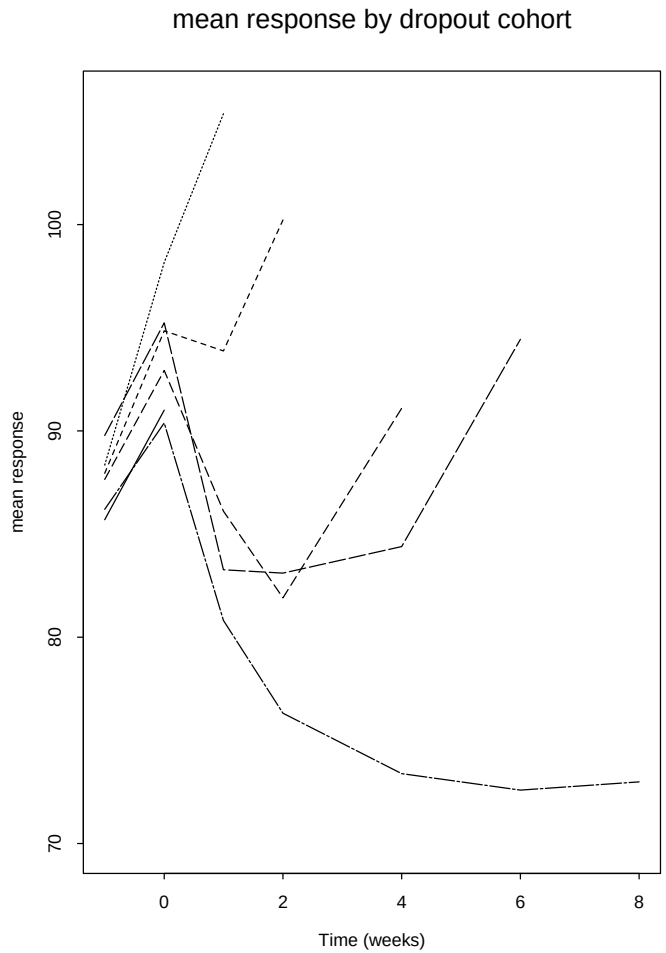
- $Z_{ij} \sim \mathrm{N}(0, \tau^2)$
- $\mu_i(t_{ij}) = X_{1i}(t_{ij})\beta_1$

Hazard sub-model

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

- $h_0(t) =$ non-parametric baseline hazard
- $\eta_2(t) = X_{2i}(t) + R_{2i}(t) =$ linear predictor for hazard

Schizophrenia trial data
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

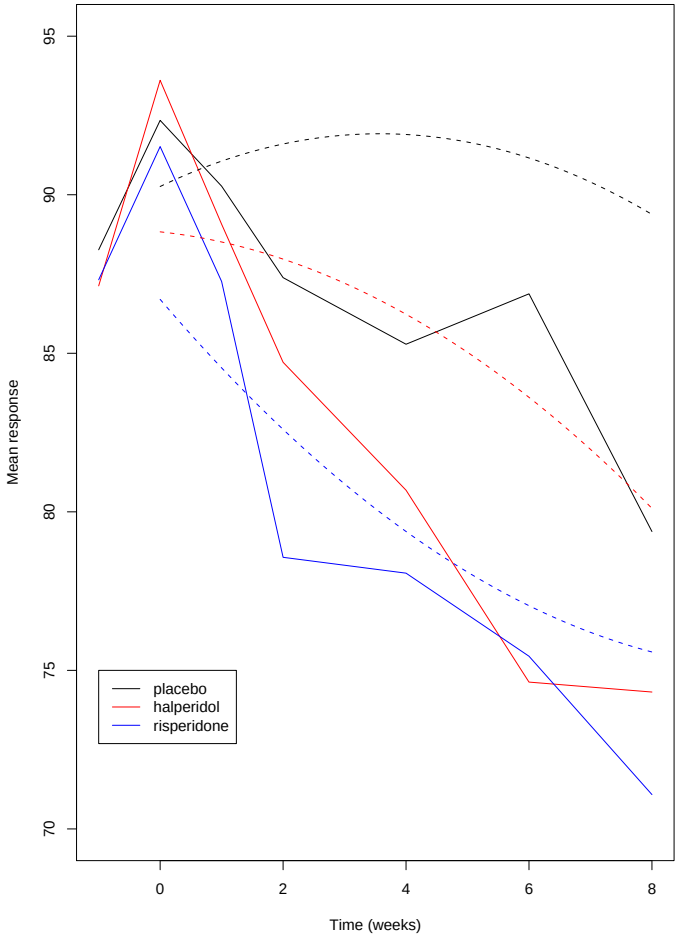
Illustrative choices for measurement process component:

$$\begin{aligned} \boldsymbol{R_1(t)} &= \boldsymbol{U_1 + W_1(t)} \\ \boldsymbol{R_1(t)} &= \boldsymbol{U_1 + U_2 t} \end{aligned}$$

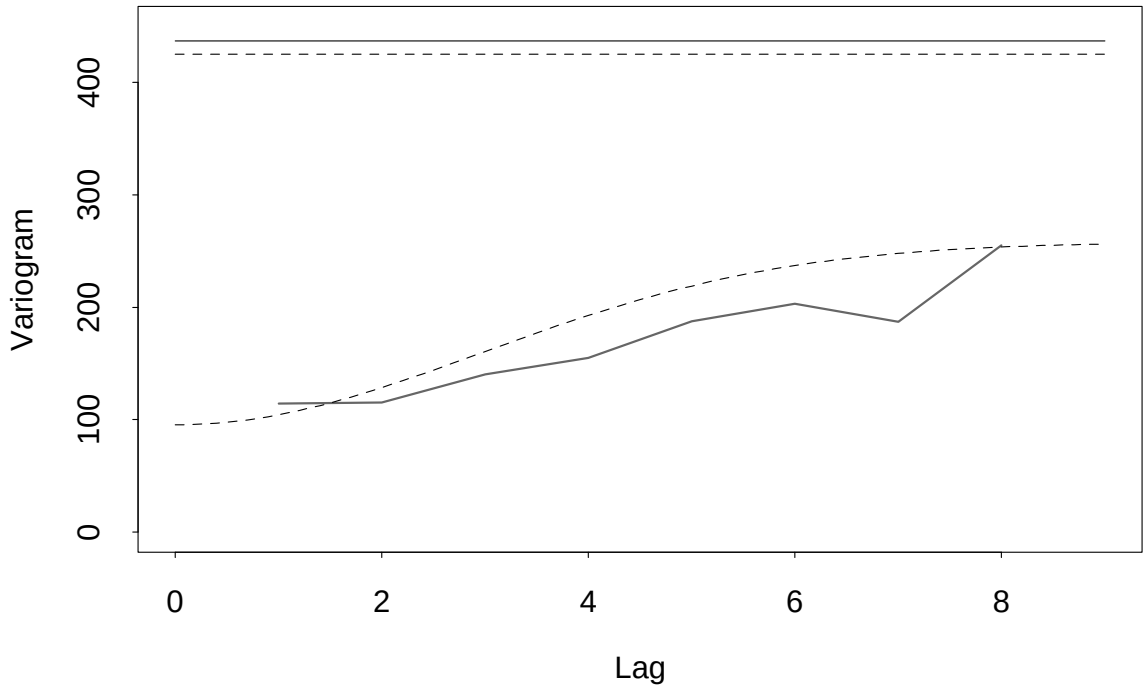
And for hazard process component:

$$\begin{aligned} \boldsymbol{R_2(t)} &= \boldsymbol{\gamma_1 R_1(t)} \\ \boldsymbol{R_2(t)} &= \boldsymbol{\gamma_1 (U_1 + U_2 t) + \gamma_2 U_2} \\ &= \boldsymbol{\gamma_1 R_1(t) + \gamma_2 U_2} \end{aligned}$$

Schizophrenia trial data
Mean response (random effects model)



Schizophrenia trial data
Empirical and fitted variograms



A simple transformation model

$$(Y, \log F) \sim \mathrm{MVN}(\mu, \Sigma)$$

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

Cox, 1999

Comparing approaches

Random effects models

- intuitively appealing
- flexible
- more-or-less essential for subject-level prediction

But

- likelihood-based inference computationally intensive
- robustness to non-Normality suspect

Transformation model

- very simple to use
- transparent diagnostic checks

But

- purely empirical
- requires more-or-less balanced data

More on the transformation model

- the likelihood function
- missing values and censoring
- modelling the covariance structure
- diagnostics

The likelihood function

- **Write $S = \log F$, hence $[Y, S] = \mathrm{MVN}(\mu, \Sigma)$**
- **Use factorisation $[Y, S] = [Y][S|Y]$**
- **$\mu = (\mu_Y, \mu_S)$**
- **Standard result for $[S|Y]$**
  - **$S|Y \sim \mathrm{N}(\mu_{S|Y}, \sigma^2_{S|Y})$**
  - **$\mu_{S|Y} = \mu_S + g'(\phi)V(\theta)^{-1}(Y - \mu_Y)$**
  - **$\sigma^2_{S|Y} = \nu^2 - g'(\phi)V(\theta)^{-1}g(\phi)$**

Missing values and censoring

- uncensored S_i:

  $[\boldsymbol{Y}_i] \times [S_i|\boldsymbol{Y}_i]$

- right-censored $S_i > t_{ij}$

  $[\boldsymbol{Y}_i] \times [1 - \Phi\{(t_{ij} - \mu_{S|\boldsymbol{Y}_i})/\sigma_{S|\boldsymbol{Y}}\}]$

- interval-censored $t_{ij} < S_i < t_{i,j+1}$

  $[\boldsymbol{Y}_i] \times [\Phi\{(t_{i,j+1} - \mu_{S|\boldsymbol{Y}_i})/\sigma_{S|\boldsymbol{Y}}\} - \Phi\{(t_{ij} - \mu_{S|\boldsymbol{Y}_i})/\sigma_{S|\boldsymbol{Y}}\}]$

- missing Y_{ij}

  - reduce dimensionality of $\boldsymbol{Y}_i$ accordingly
  - OK for Y_{ij} intermittently missing and/or Y_{ij} missing because $S_i < \log t_{ij}$

Modelling the covariance structure

- **Notation for covariance structure:**
  - $\mathbf{Var}(Y) = V(\theta)$
  - $\mathbf{Var}(S) = \nu^2$
  - $g(\phi) = \mathbf{Cov}(Y, S)$
- **Standard choices for $V(\theta)$ include:**
  - **Random intercept and slope (Laird and Ware, 1982)**

    $$Y_{ij} - \mu_{ij} = A_i + B_i t_{ij} + Z_{ij} : j = 1, .., n_i; i = 1, ..., m$$

  - **Three components of variation (Diggle, 1988)**

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

  - **Compound symmetry**

    $$Yij - \mu_{ij} = A_i + Z_{ij}$$

- **Models for $g(\phi)$?**
  - **uniform correlation**
  - **saturated**
  - **intermediate?**

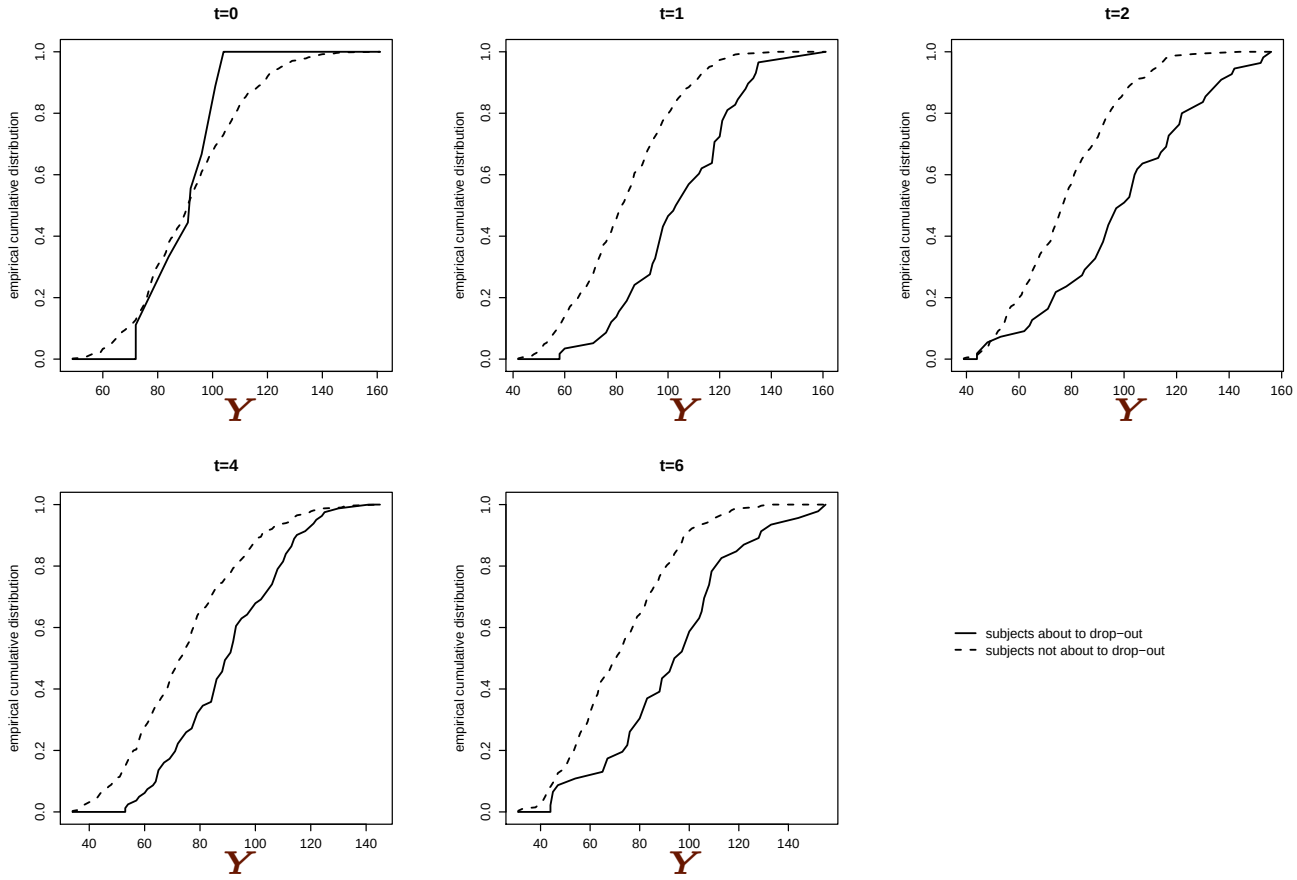
  Choice for $V(\theta)$ implies constraints on $g(\phi)$

Diagnostics

Assume balanced data, i.e. $t_{ij} = t_j$

- **Fit to $[Y]$:**
  - **consider all 'survivors" at each follow-up time t_j**
  - **classify according to whether they do or do not survive to time t_{j+1}**
  - **check goodness-of-fit to distributions implied by the model**
- **Fit to $[S|Y]$:**
  - **Gaussian P-P and Q-Q plots with multiple imputation of censored $\log S$**
  - **Check that deviation from linearity is comparable with simulated $\mathrm{N}(0, 1)$ samples.**

Re-analysis of schizophrenia trial data
Dropout is not completely at random



Re-analysis of schizophrenia trial data
Model specification

- measurements, Y: random intercept and slope

$$Y_{ij} - \mu_{ij} = A_i + B_i t_{ij} + Z_{ij} : j = 1, .., n_i; i = 1, ..., m$$

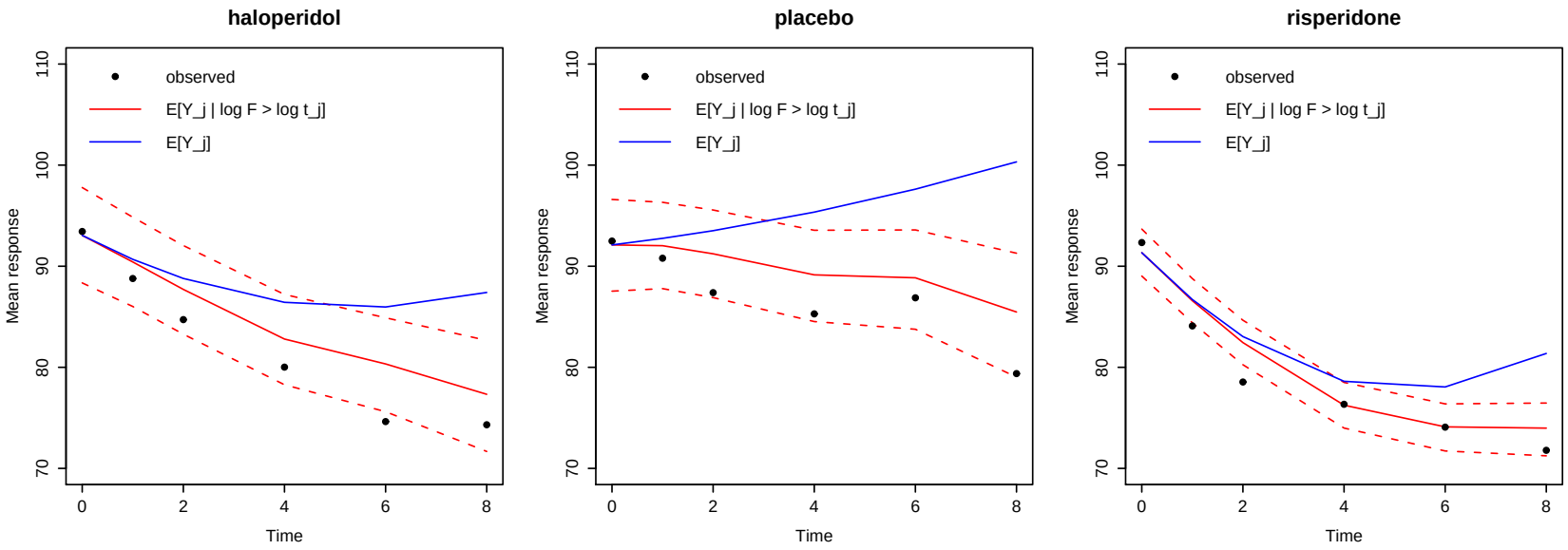
- dropout time, F

$$S = \log F \sim \mathrm{N}(\mu_S, \nu^2)$$

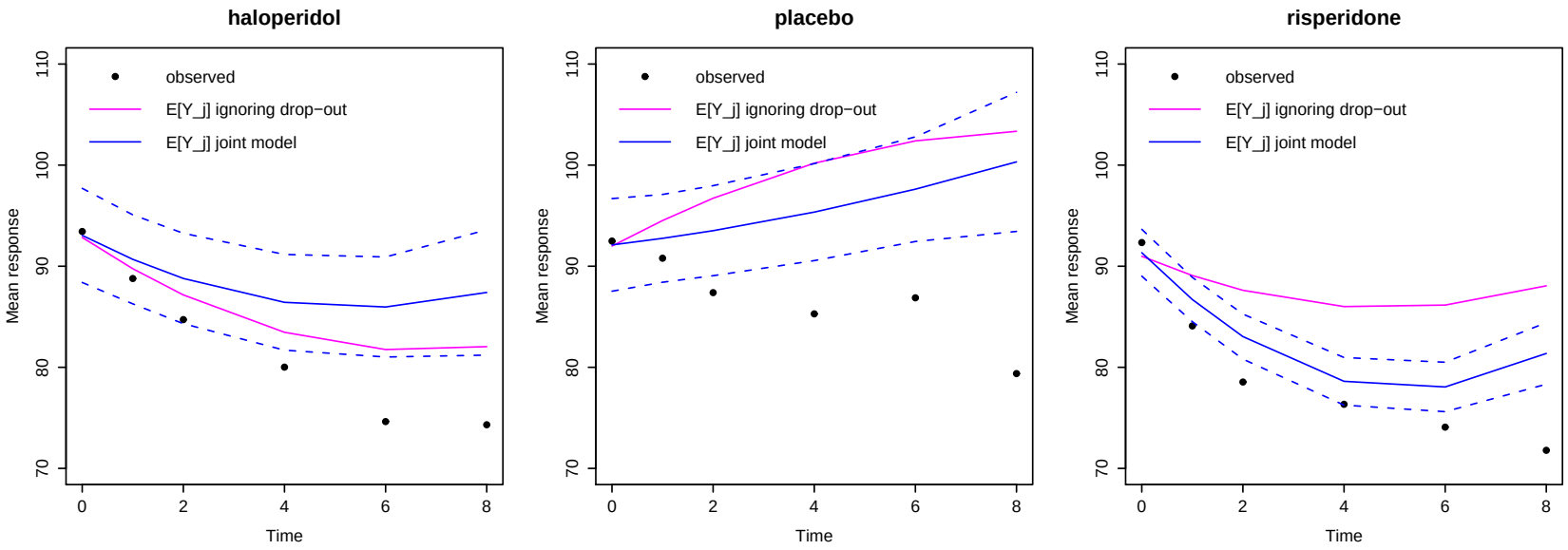
- cross-covariances

$$\mathrm{Cov}(Y_j, S) = \phi_j : j = 1, ..., 6$$

Re-analysis of schizophrenia trial data
Goodness-of-fit: mean response profiles



Re-analysis of schizophrenia trial data
Fitted mean response profiles



Closing remarks

- the role of modelling

  "We buy information with assumptions"

  Coombs (1964)

- choice of model/method should relate to scientific purpose.

  "Analyse problems, not data"

  PJD

- simple models/methods are useful when exploring a range of modelling options, for example to select from many potential covariates.

- **complex models/methods are useful when seeking to understand subject-level stochastic variation.**
- **likelihood-based inference is usually a good idea**
- **different models may fit a data-set almost equally well**
- **`joineR` library under development**
- **longitudinal analysis is challenging, but rewarding**

> **"La peinture de l'huile,**
> **c'est tres difficile**
> **Mais c'est beaucoup plus beau,**
> **que la peinture de l'eau"**

Winston Churchill

Reading list

Books

The course is based on selected chapters from Diggle, Heagerty, Liang and Zeger (2002). Fitzmaurice, Laird and Ware (2004) covers similar ground. Fitzmaurice, Davidian, Verbeke and Molenberghs (2009) is an extensive edited compilation. Verbke and Molenberghs (2000) and Molenberghs and Verbeke (2005) are companion volumes that together cover linear models for continuous data and a range of models for discrete data. Andersen, Borgan, Gill and Keiding (1993) is a detailed account of modern methods of survival analysis and related topics. Daniels and Hogan (2008) covers missing value methods in more detail than do the more general texts.

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

Daniels, M.J. and Hogan, J.W. (2008). *Missing Data in Longitudinal Studies*. Chapman and Hall/CRC Press

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

Fitzmaurice, G.M., Davidian, M., Verbeke, G. and Molenberghs, G. (2009). *Longitudinal Data Analysis*. Boca Raton: Chapman and Hall/CRC

Fitzmaurice, G.M., Laird, N.M. and Ware, J.H. (2004). *Applied Longitudinal Analysis*. New Jersey: Wiley.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. New York: Springer

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer.

Journal articles

Berzuini, C. and Larizza, C. (1996). A unified approach for modelling longitudinal and failure time data, with application in medical monitoring. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **18**, 109-123.

Brown E.R. and Ibrahim, J.G. (2003). A Bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics*, **59**, 221-228.

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. , Farewell, D. and Henderson, R. (2007). Longitudinal data with dropout: objectives, assumptions and a proposal (with Discussion). *Applied Statistics*, **56**, 499–550. Diggle, P.J., Heagerty, P., Liang, K-Y and Zeger, S.L. (2002). *Analysis of Longitudinal Data (second edition)*. Oxford : Oxford University Press.

Dobson, A. and Henderson, R. (2003). Diagnostics for joint longitudinal and dropout time modelling. *Biometrics*, **59**, 741-751.

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.

Lim, E., Ali, A., Theodorou, P., Sousa, I., Ashrafian, H., Chamageorgakis, A.D., Henein, M., Diggle, P. and Pepper, J, (2008). A longitudinal study of the profile and predictors of left ventricular mass regression after stentless aortic valve replacement. *Annals of Thoracic Surgery*, **85**, 2026–2029.

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.

Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. American Statistical Association*, **90**, 106-121.

Rotnitzky, A., Robins, J.M. and Scharfstein, D.O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *J. American Statistical Association*, **93**, 1321-1339.

Scharfstein, D.O., Rotnitzky, A. and Robins, J.M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J. American Statistical Association*, **94**, 1096-1146.

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. and Davidian, M.A. (2004). Joint modelling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, **14**, 809-834.

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.