Longitudinal Data Analysis

Peter Diggle

(Lancaster University and University of Liverpool)

Belfast, February 2015

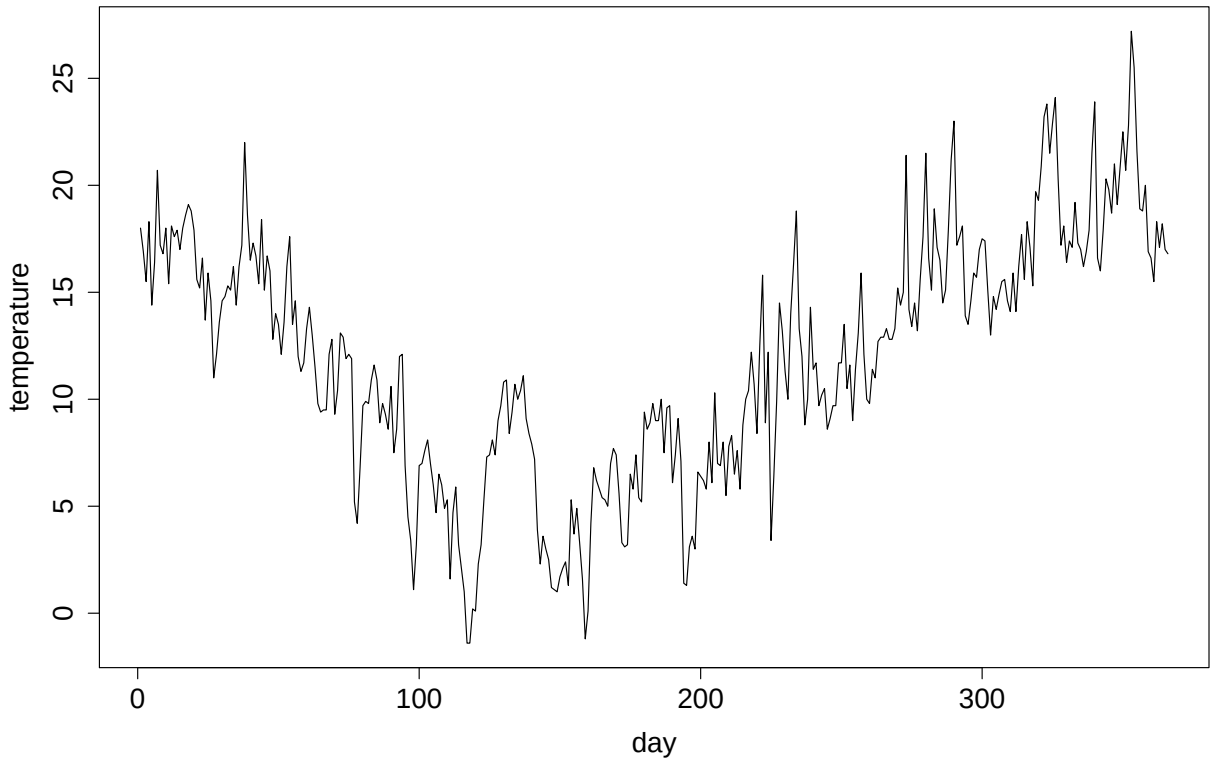
Lecture topics

1. **Time series and longitudinal data: similarities/differences**
2. **Linear models: capturing correlation structure**
3. **Missing values: Rubin's hierarchy, informative dropout**
4. **Generalised linear models: binary and count data**
5. **Joint modelling: repeated measurement and time-to-event outcomes**

1. Time series and longitudinal data

Bailrigg temperature records

Daily maximum temperatures, 1.09.1995 to 31.08.1996

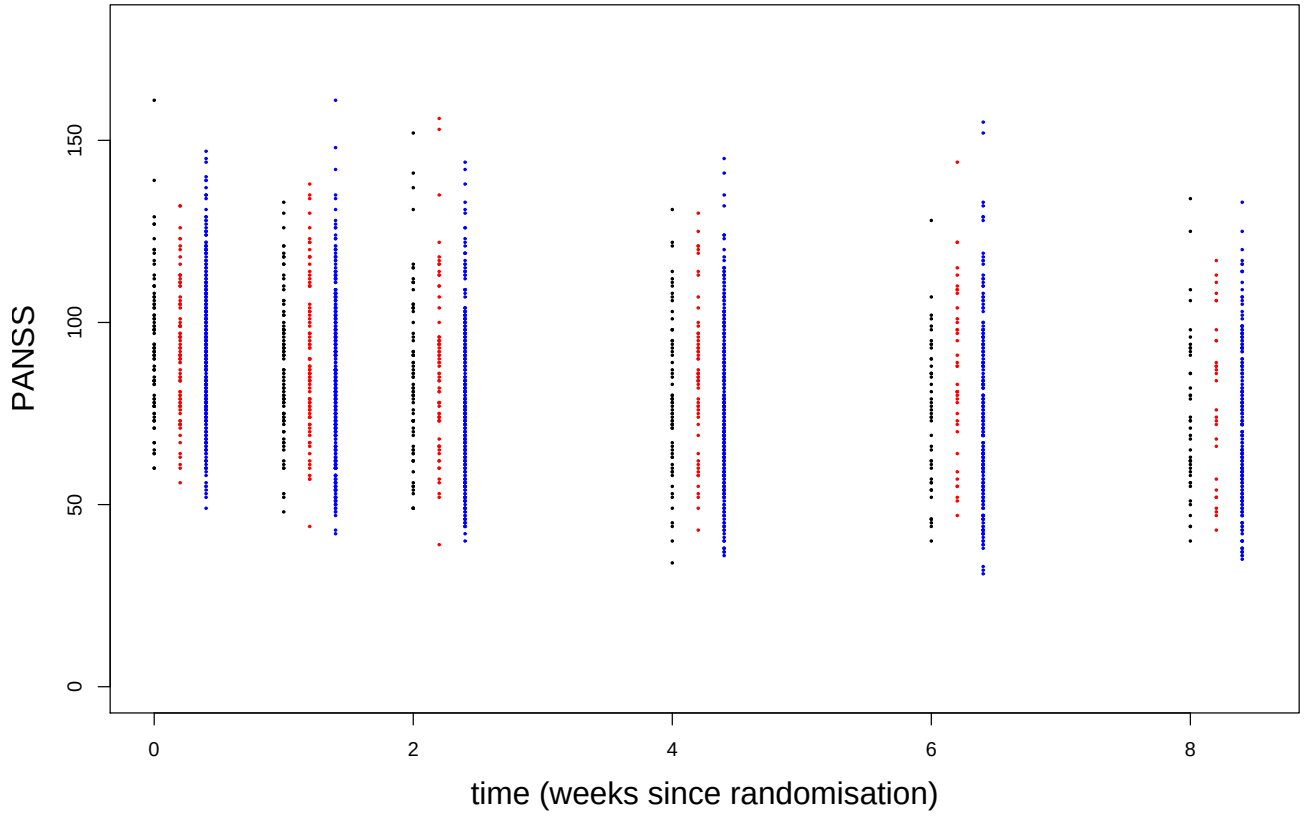


Schizophrenia clinical trial (PANSS)

- 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



Diggle, Farewell and Henderson (2007)

Time series decomposition

- trend and residual
- autocorrelation
- prediction

Analysis of Bailrigg temperature data

```
data<-read.table("../data_and_figures/maxtemp.data",header=F)
temperature<-data[,4]
n<-length(temperature)
day<-1:n
plot(day,temperature,type="l",cex.lab=1.5,cex.axis=1.5)
#
# plot shows strong seasonal variation,
# try simple harmonic regression
#
```

```
c1<-cos(2*pi*day/n)
s1<-sin(2*pi*day/n)
fit1<-lm(temperature~c1+s1)
lines(day,fit1$fitted.values,col="red")
#
# add first harmonic of annual frequency to check for
# non-sinusoidal pattern
#
c2<-cos(4*pi*day/n)
s2<-sin(4*pi*day/n)
fit2<-lm(temperature~c1+s1+c2+s2)
lines(day,fit2$fitted.values,col="blue")
#
# two fits look similar, but conventional F test says otherwise
#
summary(fit2)
RSS1<-sum(fit1$resid^2); RSS2<-sum(fit2$resid^2)
F<-((RSS1-RSS2)/2)/(RSS2/361)
1-pf(F,2,361)
```

```
#
# conventional residual plots
#
#   residuals vs fitted values
#
plot(fit2$fitted.values,fit2$resid)
#
#   residuals in time-order as scatterplot
#
plot(1:365,fit2$resid)
#
#   and as line-graph
#
plot(1:365,fit2$resid,type="l")
```

```
#
# examine autocorrelation properties of residuals
#
residuals<-fit2$resid
par(mfrow=c(2,2),pty="s")
for (k in 1:4) {
   plot(residuals[1:(n-k)],residuals[(k+1):n],
         pch=19,cex=0.5,xlab=" ",ylab=" ",main=k)
   }
par(mfrow=c(1,1))
acf(residuals)
#
# exponentially decaying correlation looks reasonable
#
cor(residuals[1:(n-1)],residuals[2:n])
Xmat<-cbind(rep(1,n),c1,s1,c2,s2)
rho<-0.01*(60:80)
profile<-AR1.profile(temperature,Xmat,rho)
```

```
#
# examine results
#
plot(rho,profile$logl,type="l",ylab="L(rho)")
Lmax<-max(profile$logl)
crit.val<-0.5*qchisq(0.95,1)
lines(c(rho[1],rho[length(rho)]),rep(Lmax-crit.val,2),lty=2)
profile
#
# Exercise: how would you now re-assess the significance of
# the second harmonic term?
```

```
#
# profile log-likelihood function follows
#
AR1.profile<-function(y,X,rho) {
   m<-length(rho)
   logl<-rep(0,m)
   n<- length(y)
   hold<-outer(1:n,1:n,"-")
   for (i in 1:m) {
      Rmat<-rho[i]^abs(hold)
      ev<-eigen(Rmat)
      logdet<-sum(log(ev$values))
      Rinv<-ev$vectors%*%diag(1/ev$values)%*%t(ev$vectors)
      betahat<-solve(t(X)%*%Rinv%*%X)%*%t(X)%*%Rinv%*%y
      residual<- y-X%*%betahat
      logl[i]<- - logdet - n*log(c(residual)%*%Rinv%*%c(residual))
      }
   max.index<-order(logl)[m]
   Rmat<-rho[max.index]^abs(hold)
      ev<-eigen(Rmat)
      logdet<-sum(log(ev$values))
      Rinv<-ev$vectors%*%diag(1/ev$values)%*%t(ev$vectors)
      betahat<-solve(t(X)%*%Rinv%*%X)%*%t(X)%*%Rinv%*%y
      residual<- y-X%*%betahat
      sigmahat<-sqrt(c(residual)%*%Rinv%*%c(residual)/n)
   list(logl=logl,rhohat=rho[max.index],sigmahat=sigmahat,betahat=betahat)
   }
```

Longitudinal data

- **replicated time series;**
- **focus of interest often on mean values;**
- **modelling and inference can and should exploit replication**

2. Linear models

- correlation and why it matters
- exploratory analysis
- linear Gaussian models

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

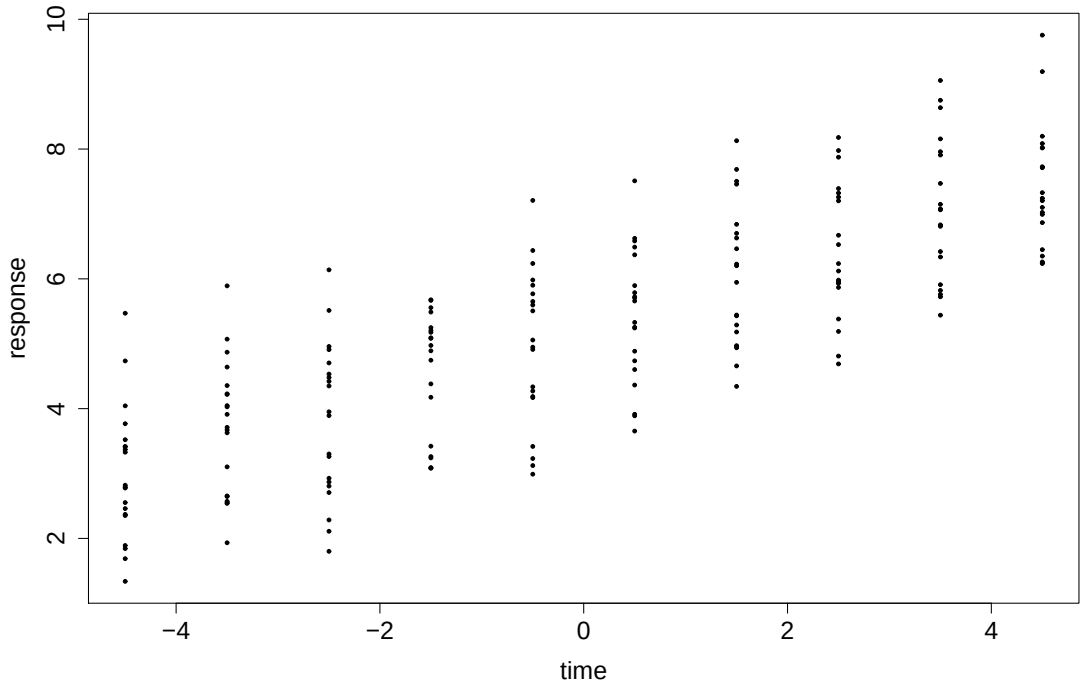
But if Y_t is a time series:

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

Exercise: is the sample variance unbiased for $\sigma^2 = \mathrm{Var}(Y_t)$?

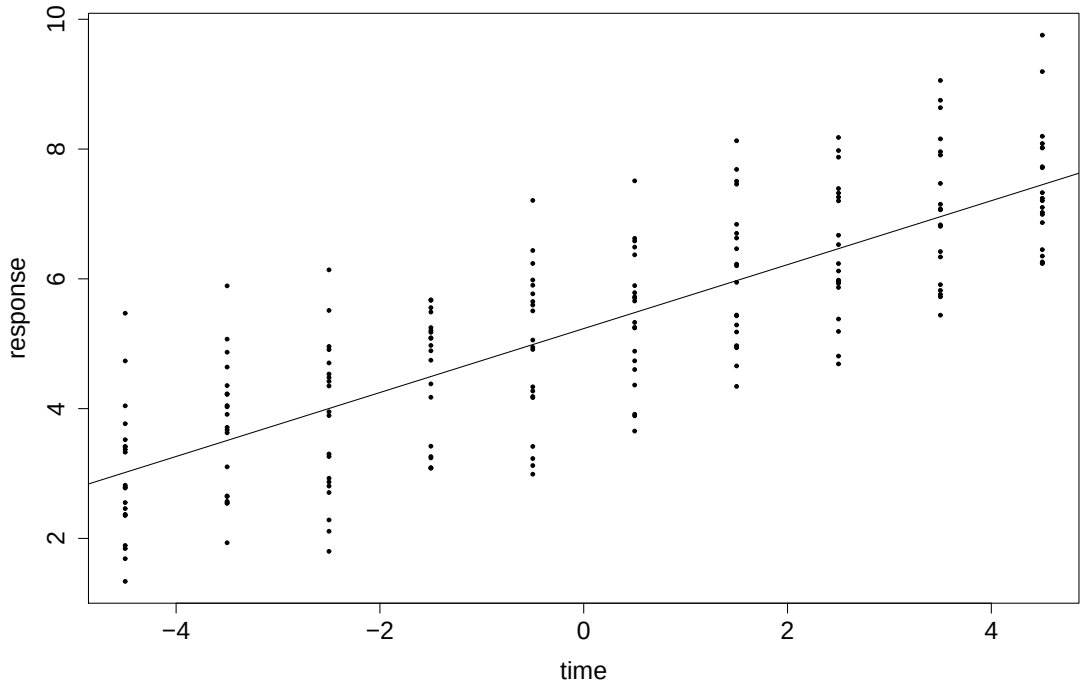
Correlation may or may not hurt you

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



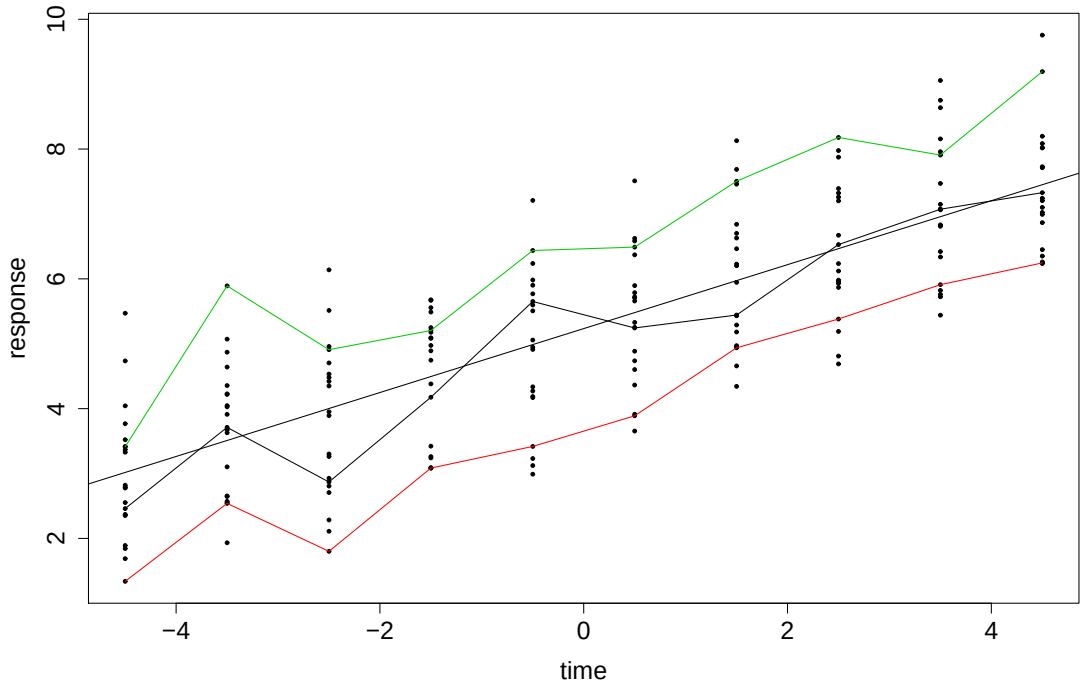
Correlation may or may not hurt you

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



Correlation may or may not hurt you

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



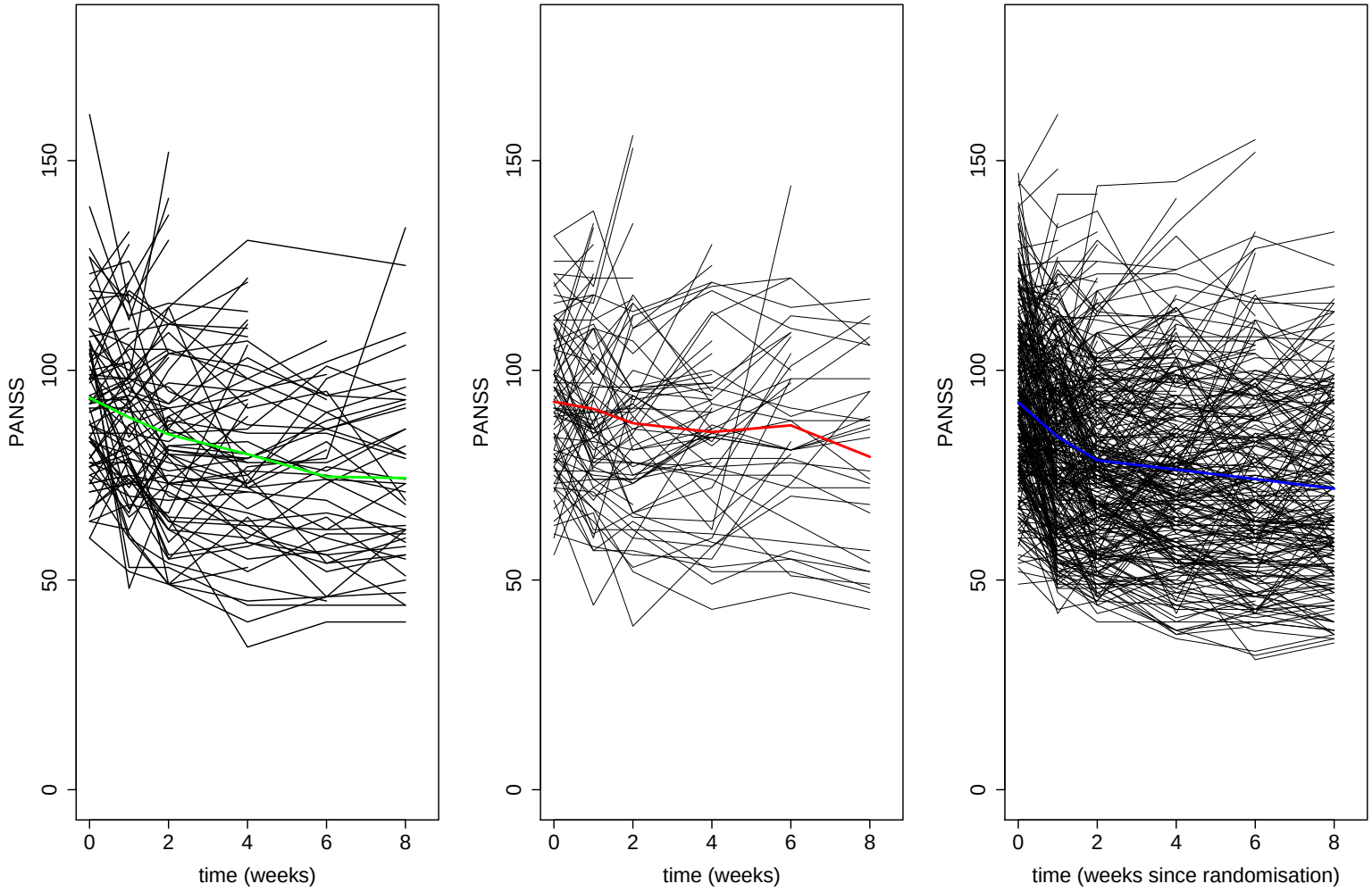
Correlation may or may not hurt you

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

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

A spaghetti plot of the PANSS data



Exploring covariance structure: balanced data

$$(Y_{ij}, t_j) : j = 1, ..., n; i = 1, ..., m$$

- fit saturated treatments-by-times model to mean response
- compute sample covariance matrix of residuals

PANSS data:

	SD	Y.t0	Y.t1	Y.t2	Y.t4	Y.t6	Y.t8
Y.t0	20.019	1.000	0.533	0.366	0.448	0.285	0.229
Y.t1	20.184	0.533	1.000	0.693	0.589	0.658	0.535
Y.t2	22.120	0.366	0.693	1.000	0.670	0.567	0.678
Y.t4	20.996	0.448	0.589	0.670	1.000	0.718	0.648
Y.t6	24.746	0.285	0.658	0.567	0.718	1.000	0.792
Y.t8	23.666	0.229	0.535	0.678	0.648	0.792	1.000

- modest increase in variability over time
- correlation decays with increasing time-separation

Exploring covariance structure: unbalanced data

$$(Y_{ij}, t_{ij}) : j = 1, ..., n_i; i = 1, ..., m$$

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,

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

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

Estimating the variogram

Data: $(Y_{ij}, t_{ij}) : i = 1, ..., m;\ j = 1, ..., n_i$

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

- Define

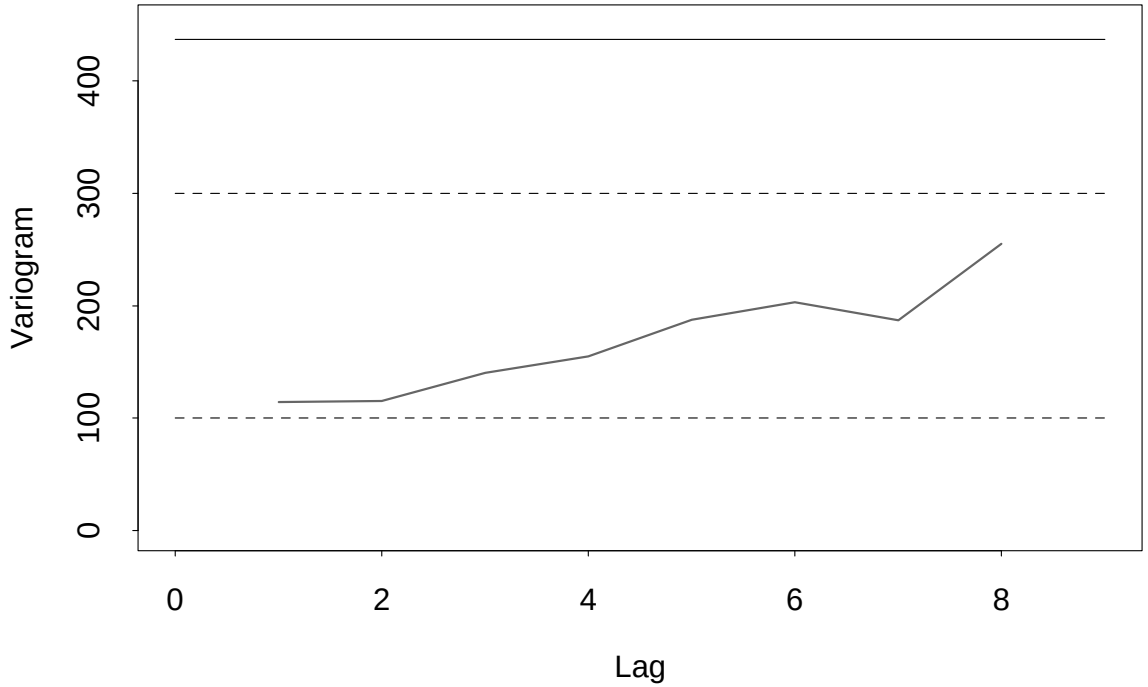
$$v_{ijk\ell} = \frac{1}{2}(r_{ij} - r_{k\ell})^2$$

- Estimate

$$\begin{aligned} \hat{V}(u) &= \text{average of all } v_{iji\ell} \text{ such that } |t_{ij} - t_{i\ell}| \simeq u \\ \hat{\sigma}^2 &= \text{average of all } v_{ijk\ell} \text{ such that } i \neq k. \end{aligned}$$

Example: sample variogram of the PANSS data

Solid lines are estimates from data, horizontal lines are eye-ball estimates (explanation later)



Where does the correlation come from?

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

General linear model, correlated residuals

i = subjects $\quad$ j = measurements within subjects

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

- 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**

Diggle, Heagerty, Liang and Zeger (2002, Chapter 5)

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**

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

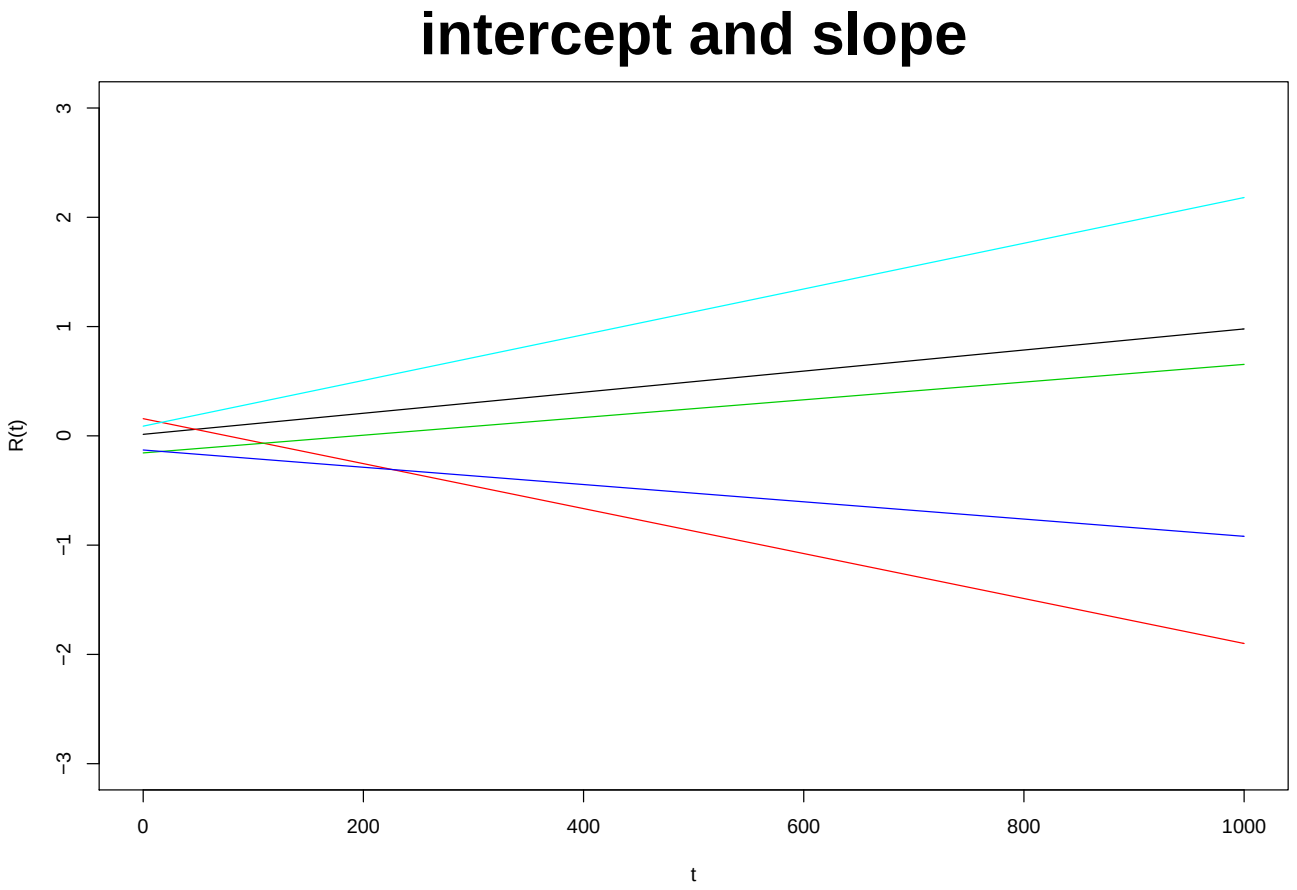
$W_i(t)$ a continuous-time Gaussian process

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

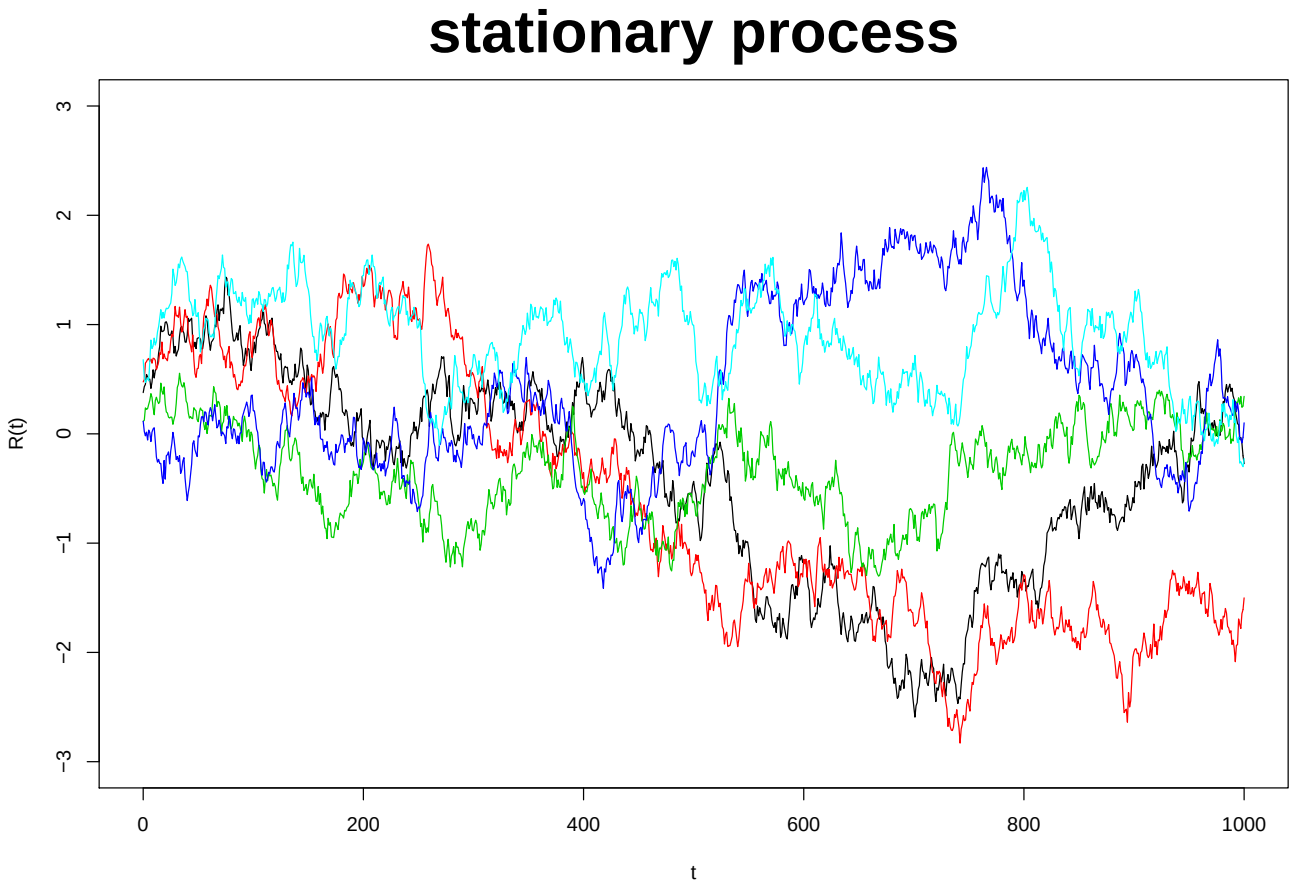
$\mathrm{Corr}\{W(t), W(t-u)\} = \rho(u)$

$\rho(u) = \exp(-u/\phi)$ 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)

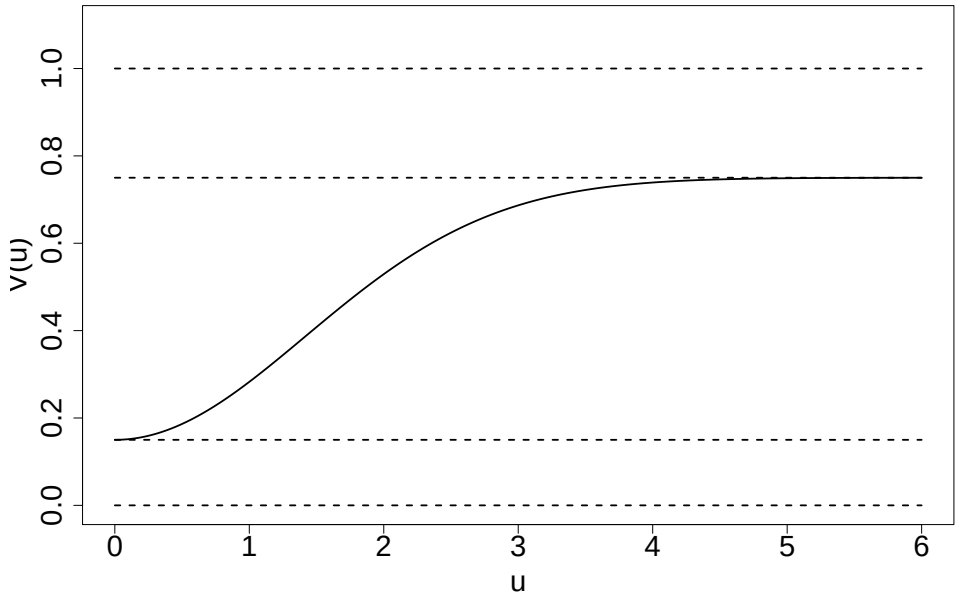
$Z_{ij} \sim \text{N}(0, \tau^2)$
(measurement errors)

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

The variogram of the general model (stationary case)

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

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



Fitting the model: 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 when there are missing values**

Maximum likelihood estimation (V_0 known)

Log-likelihood for observed data y is

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

$I \otimes V_0$ denotes block-diagonal matrix with non-zero blocks V_0

Given V_0, estimator for β is

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

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

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

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

- **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.**

A random effects model for CD4 cell counts

```
data<-read.table("../data_and_figures/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)
   }
```

```
# Simple linear model assuming uncorrelated residuals
#
fit1<-lm(CD4~time)
summary(fit1)
#
# random intercept and slope model
#
library(nlme)
?lme
fit2<-lme(CD4~time,random=~1|id)
summary(fit2)
```

```
# make fitted value constant before sero-conversion
#
timeplus<-time*(time>0)
fit3<-lme(CD4~timeplus,random=~1|id)
summary(fit3)
tfit<-0.1*(0:50)
Xfit<-cbind(rep(1,51),tfit)
fit<-c(Xfit%*%fit3$coef$fixed)
Vmat<-fit3$varFix
Vfit<-diag(Xfit%*%Vmat%*%t(Xfit))
upper<-fit+2*sqrt(Vfit)
lower<-fit-2*sqrt(Vfit)
#
# plot fit with 95% point-wise confidence intervals
#
plot(time,CD4,pch=19,cex=0.25)
lines(c(-3,tfit),c(upper[1],upper),col="red")
lines(c(-3,tfit),c(lower[1],lower),col="red")
```

3. 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 structure of longitudinal data can sometimes be exploited to good effect, by modelling how the probability of dropout for each person depends on their previously observed measurements

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**
- **MNAR (not at random): conditional on observed measurements, P(missing) depends on unobserved measurements.**

Rubin (1976)

Dropout

Once a subject goes missing, they never return

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.**

Little (1995)

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**

Proof: Partition Y for each subject into observed and missing components, $Y = (Y_o, Y_m)$ and let M denote binary vector of missingness indicators. Likelihood for observed data is

$$\begin{aligned} L = \quad g(y_o, m) &= \int f(y_o, y_m, m) dy_m \\ &= \int f(y_o) f(y_m|y_o) p(m|y_o, y_m) dy_m \end{aligned}$$

If $p(m|y_o, y_m) = p(m|y_o)$, take outside integral to give

$$L = p(m|y_o) f(y_o)$$

and log-likelihood contribution

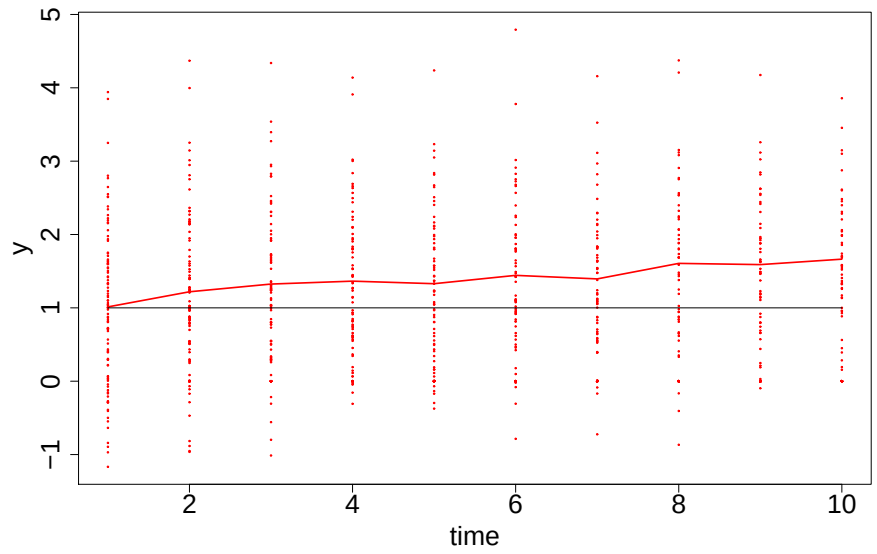
$$\log L = \log p(m|y_o; \theta) + \log f(y_o|\theta)$$

- OK to ignore first term for likelihood inference about θ
- and no loss of efficiency if $\theta = (\theta_1, \theta_2)$ such that θ_1 and θ_2 parameterise $p(\cdot)$ and $f(\cdot)$, respectively.

But is inference about $f(\cdot)$ what you want?

Example

- **Model is $Y_{ij} = \mu + U_i + Z_{ij}$ (random intercept)**
- **Dropout is MAR: $\text{logit}(p_{ij}) = -1 - 2 \times Y_{i,j-1}$**



- **Observed means increase over time, but population mean μ is constant**

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.**

Diggle, Farewell and Henderson (2007)

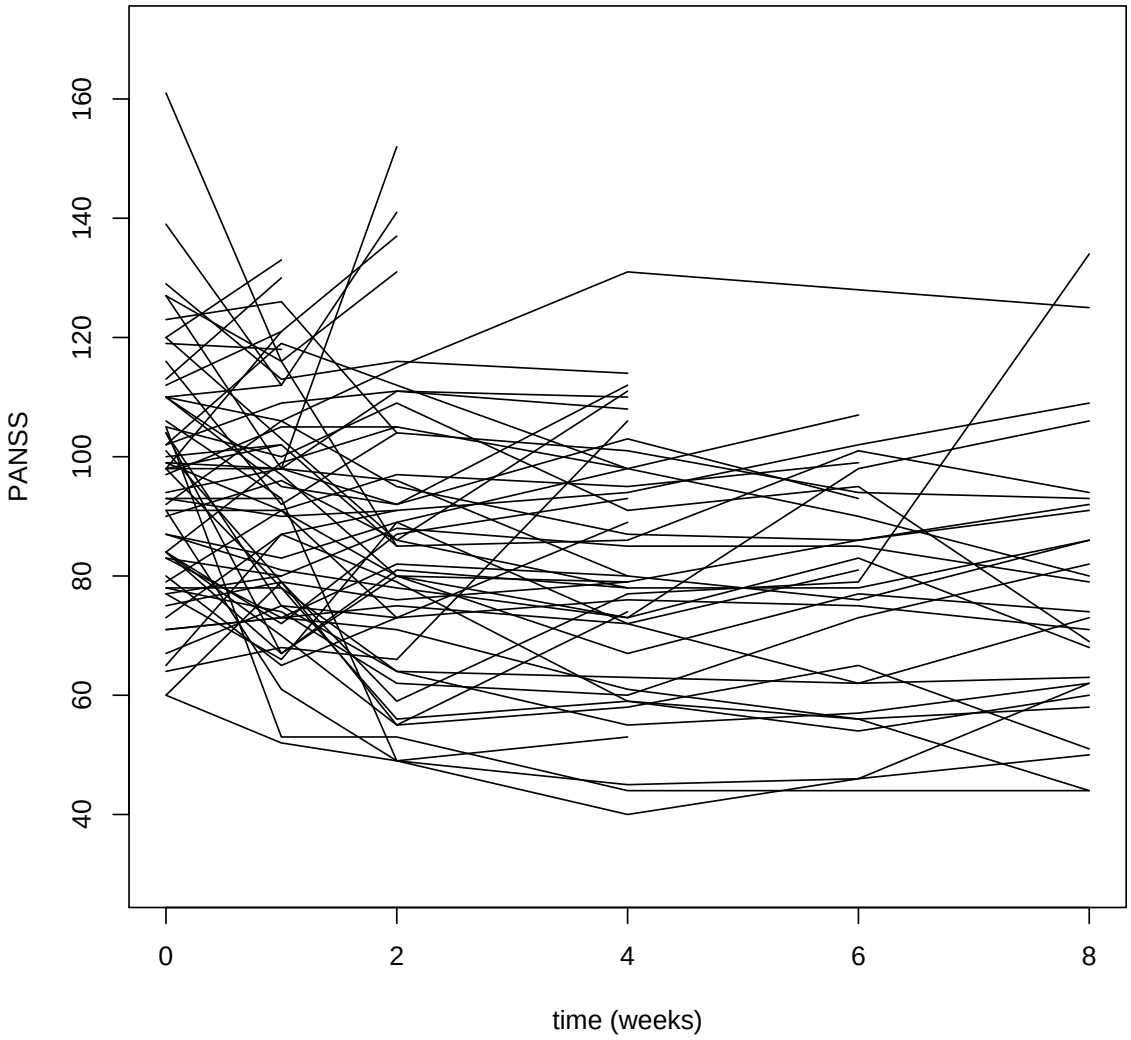
Schizophrenia trial data

- 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:

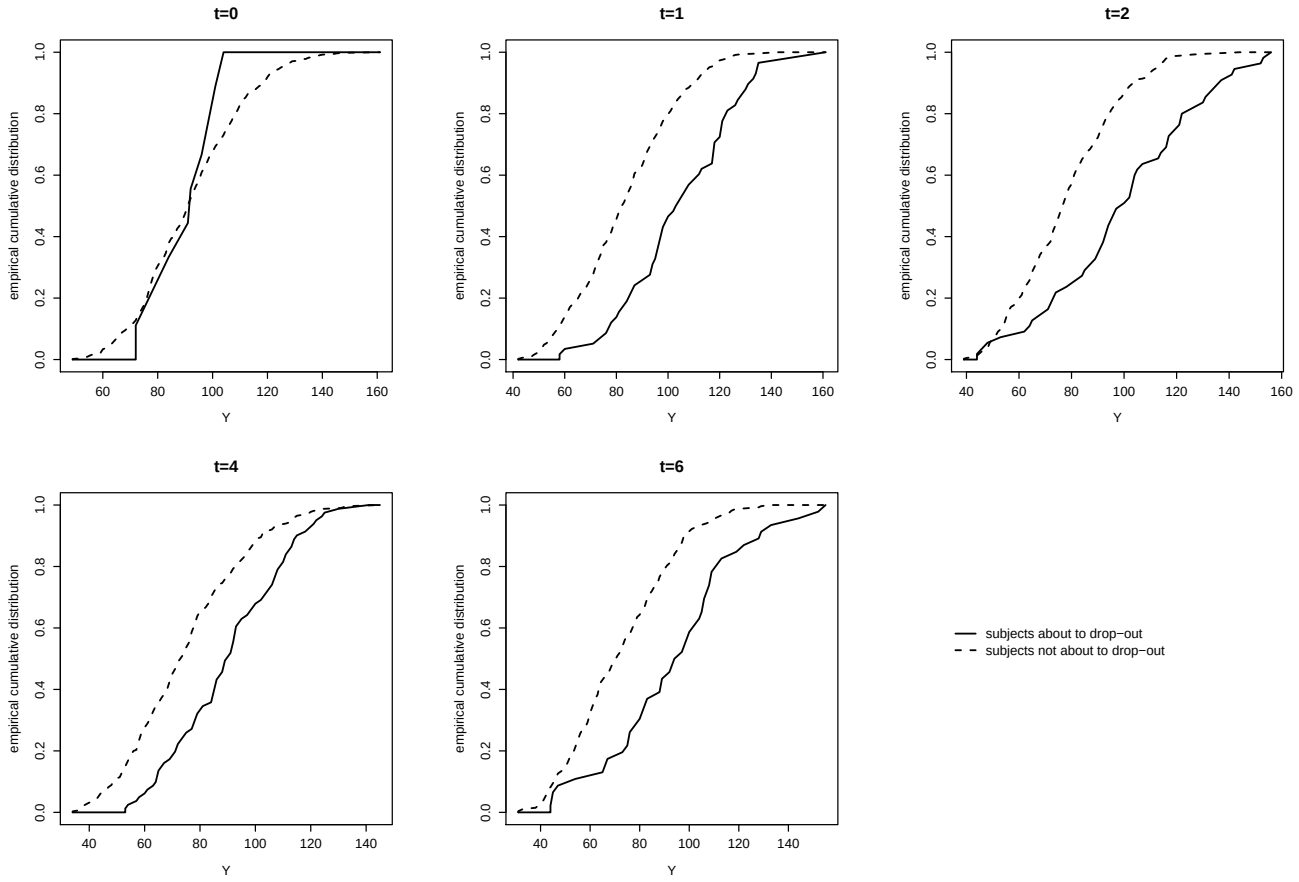
week	−1	0	1	2	4	6	8
missing	0	3	9	70	122	205	251
proportion	0.00	0.01	0.02	0.14	0.24	0.40	0.49

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

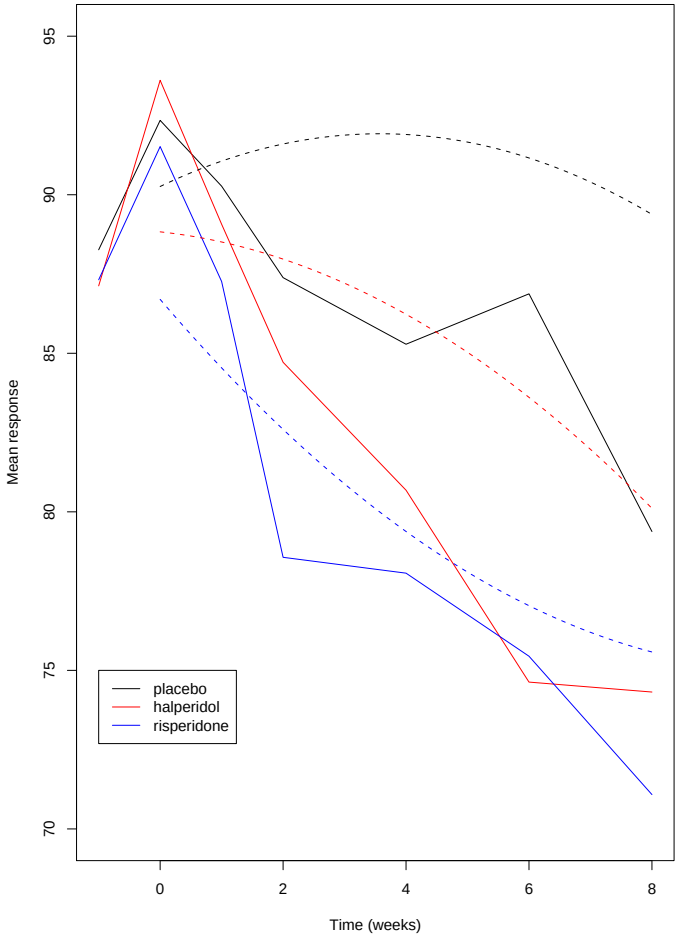
Schizophrenia data
PANSS responses from haloperidol arm



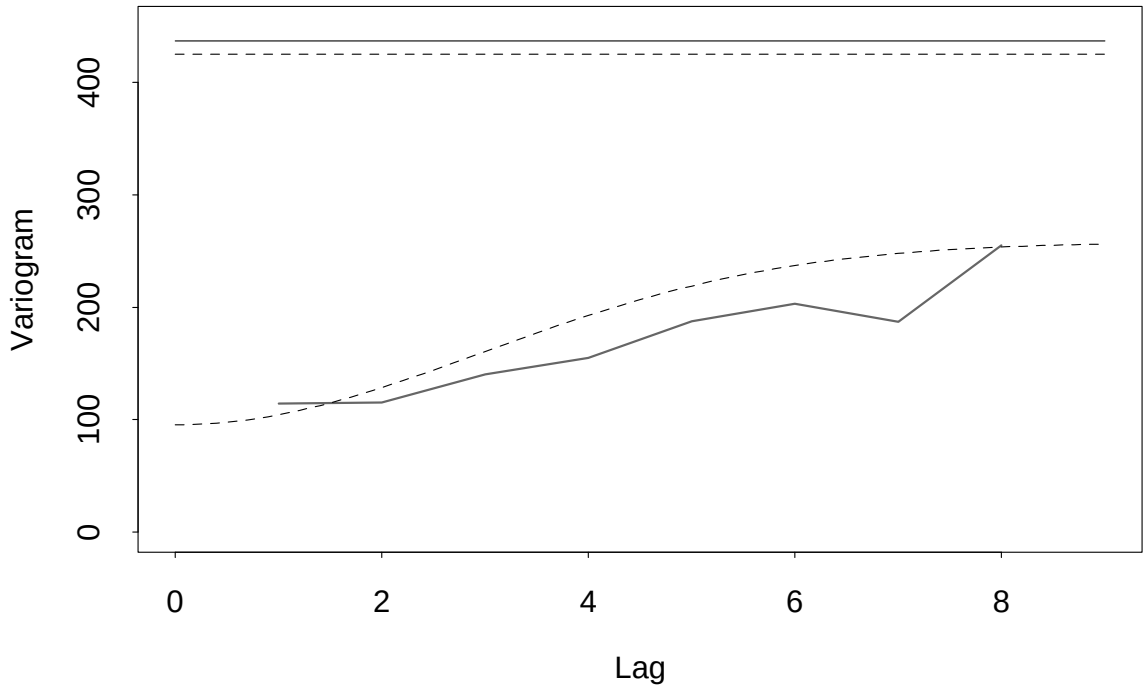
Dropout is not completely at random



Schizophrenia trial data
Mean response (random effects model)



Schizophrenia trial data
Empirical and fitted variograms



Schizophrenia trial data: summary

- dropout is not MCAR
- MAR model apparently fits well, but:
  - hard to distinguish empirically between different MAR models;
  - and we haven't formally investigated evidence for informative dropout
- Fitted means relate to hypothetical, dropout-free population

Embedding MAR within an MNAR model

1. **Diggle and Kenward**

   Measurement model

   General linear model for $Y_i = \{Y_{it} : t = 1, ..., n\}$ (balanced data)

   Dropout model

   Logistic regression:

$$\text{logit } \mathrm{P}(D_i = t | Y_i) = \alpha + \beta Y_{i,t-1} + \gamma Y_{it}$$

Diggle and Kenward, 1994

2. Barrett, Diggle, Henderson and Taylor-Robinson

Hybrid time-scales

Continuous-time measurement model

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

Discrete-time survival model

$$U_{ik} = \{S_i(u_k) : k = 1, ..., N\}$$

Linkage

$$\mathrm{P}(D_i = d | D_i > d-1, U) = 1 - \Phi\left(\tilde{\mu}_{id} + \sum_{k=1}^{d} \gamma_k U_{ik}\right)$$

Barrett, Diggle, Henderson and Taylor-Robinson, 2015

4. Generalized linear models

- random effects models
- transition models
- marginal models

Diggle, Heagerty, Liang and Zeger (2002, Chapter 7)

Random effects GLM

Responses $Y_1, \ldots, Y_n$ on an individual subject conditionally independent, given unobserved vector of random effects U

$U \sim g(u)$ represents properties of individual subjects that vary randomly between subjects

- $E(Y_j|U) = \mu_j : h(\mu_i) = \mathbf{x}_j'\beta + U'\alpha$
- $\mathrm{Var}(Y_j|U) = \phi v(\mu_j)$
- $(Y_1, \ldots, Y_n)$ are mutually independent conditional on U.

Likelihood inference requires evaluation of

$$f(y) = \int \prod_{j=1}^{n} f(y_j|U)g(U)dU$$

Transition GLM

Each Y_j modelled conditionally on preceding $Y_1, Y_2, \ldots, Y_{j-1}$.

- $E(Y_j|\text{history}) = \mu_j$
- $h(\mu_j) = \mathbf{x}_j'\beta + \sum_{k=1}^{j-1} \mathbf{Y}_{j-k}'\alpha_k$
- $\text{Var}(Y_j|\text{history}) = \phi v(\mu_j)$

Construct likelihood as product of conditional distributions, usually assuming restricted form of dependence.

Example: $f_k(y_j|y_1, ..., y_{j-1}) = f_k(y_j|y_{j-1})$

Need to condition on y_1 as model does not directly specify marginal distribution $f_1(y_1)$.

Marginal GLM

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

- $E(y) = \mu : h(\mu) = X\beta$
- $\text{Var}(y_i) = \phi v(\mu_i)$
- $\text{Corr}(y) = R(\alpha)$.

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 generalized estimating equations

Liang and Zeger (1986)

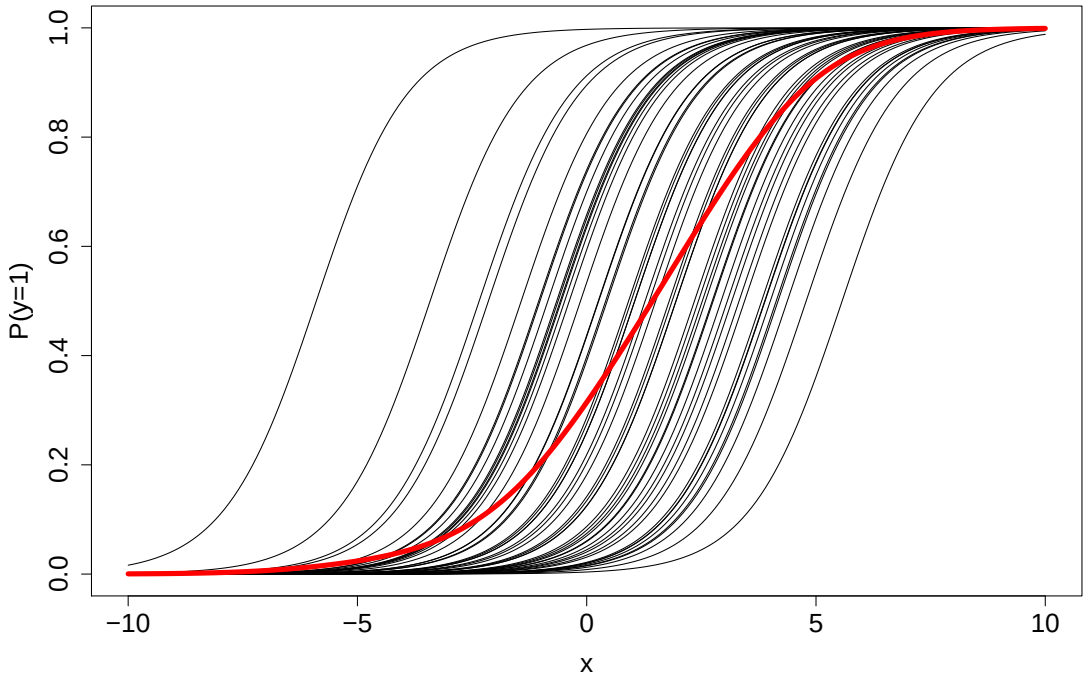
What are we estimating?

- 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

R software

The following is almost certainly an incomplete list.

- **Marginal models**

  Function `gee` within package of same name

- **Random effects models**

  Function `glmmPQL` within `MASS` package or `lmer` within `lme4` (but note evaluation of likelihood uses approximate methods that may perform badly if random effects are high-dimensional). Package `glmmBUGS` is a Bayesian alternative.

- **Transition models**

  Standard `glm` function, after computing values of required functions of lagged responses to be used as explanatory variables.

Illustration of marginal modelling

```
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)
```

5. Joint modelling: repeated measurements and time-to-event outcomes

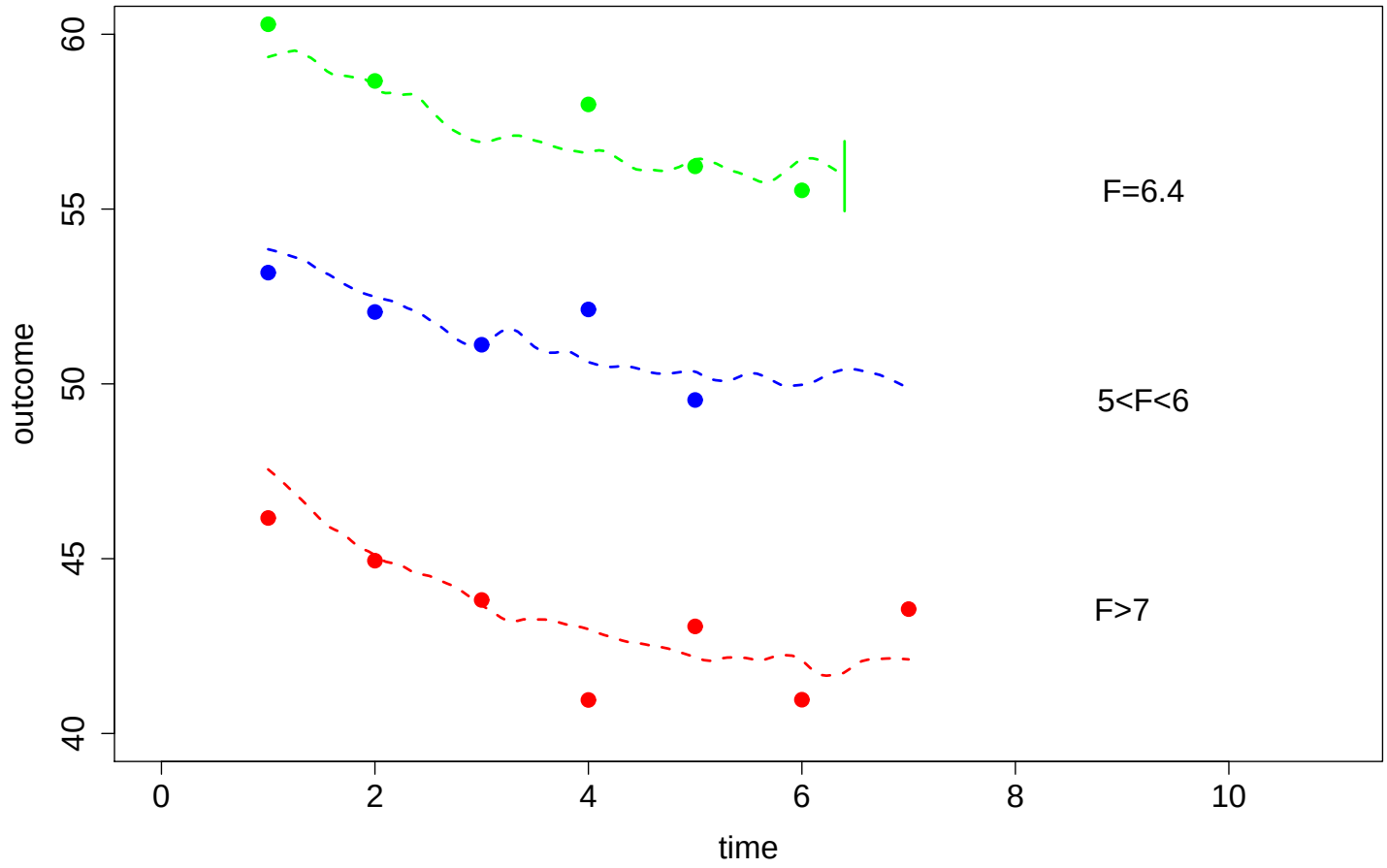
- what is it?
- why do it?
- random effects models

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.

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

Joint modelling: what is it?



Joint modelling: why do it?

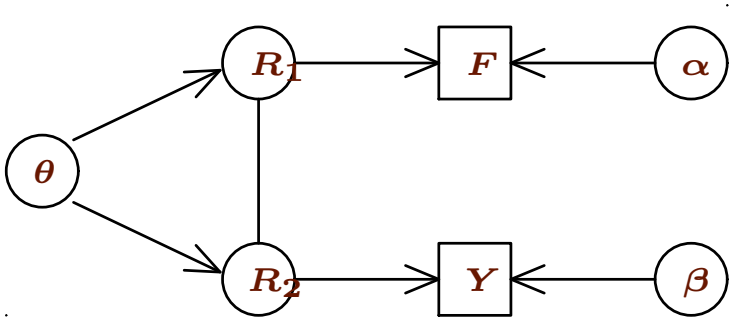
- To analyse failure time F, whilst exploiting correlation with
  an imperfectly measured, time-varying risk-factor Y

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

- Because relationship between Y and F is of direct interest

Random effects models

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



Example: Wulfsohn and Tsiatis, 1997

latent random effect; measurement model; hazard model

Latent random effect

Random intercept and slope: $U_i = (U_{0i}, U_{1i})$

Laird and Ware, 1982

Measurement model

$$Y_{ij} = \mu_i(t_{ij}) + U_{0i} + U_{1i}t_{ij} + Z_{ij}$$

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

Hazard model

$$h_i(t) = h_0(t) \exp\{\theta(U_{0i} + U_{1i} t_{ij})\}$$

- $h_0(t)$ = non-parametric baseline hazard
- $\theta(U_{0i} + U_{1i} t_{ij})$ = linear predictor for hazard, proportional to random effect

Example: Henderson, Diggle and Dobson, 2000

latent stochastic process; measurement model; hazard 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 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 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

Two (relatively) open questions

- **Repeated measurements and recurrent events**
- **Informative follow-up**

Note: what constitutes a missing value if follow-up schedule is not pre-specified?

The joineR package

Exploring the mental data-set

```
library(joineR)
data(mental)
mental[1:5,]
y<-as.matrix(mental[,2:7]) # convert data to matrix format
means<-matrix(0,3,6)
for (trt in 1:3) {
   ysub<-y[mental$treat==trt,]
   means[trt,]<-apply(ysub,2,mean,na.rm=TRUE)
   }
residuals<-matrix(0,150,6)
for (i in 1:150) {
   residuals[i,]<-y[i,]-means[mental$treat[i],]
   }
V<-cov(residuals,use="pairwise"); R<-cor(residuals,use="pairwise")
round(cbind(diag(V),R),3)
```

The joineR package

Setting up a jointdata object

```
is.data.frame(mental)
mental.unbalanced<-to.unbalanced(mental, id.col = 1,
   times = c(0,1,2,4,6,8),Y.col = 2:7, other.col = 8:11)
names(mental.unbalanced)
names(mental.unbalanced)[3]<-"Y"
mental.long<-mental.unbalanced[,1:3]
mental.surv <- UniqueVariables(mental.unbalanced,
   var.col=6:7,id.col = 1)
mental.baseline <- UniqueVariables(mental.unbalanced,
   var.col=4,id.col = 1)
mental.baseline$treat<-as.factor(mental.baseline$treat) # !!!
mental.joint<-jointdata(longitudinal=mental.long,
   survival=mental.surv,baseline=mental.baseline,
   id.col="id",time.col="time")
summary(mental.joint)
```

The joineR package

Fitting a joint model

```
fit<-joint(mental.joint,long.formula=Y~-1+treat+time,
          surv.formula=Surv(surv.time,cens.ind)~treat,
          model="intslope")
summary(fit)
set.seed(389712)
fit.se <- jointSE(fit, n.boot = 5)
# use much larger number of bootstrap samples in practice
set.seed(54912)
fit.se100 <- jointSE(fit, n.boot = 100,max.it=2000,tol=0.01,
       print.detail=TRUE)
fit.se100
```

Take-home messages

- **Correlation matters**
- **Longitudinal designs address a richer set of questions than cross-sectional designs**
- **But also raise challenges in formulating a valid, efficient analysis:**
  - **what, precisely, is the question?**
  - **what explicit and implicit assumptions does the proposed method of analysis make?**

References

Barrett, J., Diggle, P.J., Henderson, R. and Taylor-Robinson, D. (2013). Joint modelling of repeated measurements and time-to-event outcomes: flexible model specification and exact likelihood inference. *Journal of the Royal Statistical Society*, B (submitted)

Chatfield, C. (2003). *The Analysis of Time Series: an Introduction (6th edition)*. London: Chapman and Hall.

Diggle, P.J. (1990). *Time Series: a Biostatistical Introduction.*

Diggle, P.J. and Al Wasel, I. (1997). Spectral analysis of replicated biomedical time series (with Discussion). *Applied Statistics*, **46**, 31–71.

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.

Diggle, P.J. and Kenward, M.G. (1994). Informative dropout in longitudinal data analysis (with Discussion). *Appl. Statist.* **43**, 49–93.

Diggle, P.J. and Sousa, I. (2013). Real-time detection of incipient renal failure in primary care patients using a dynamic time series model. *Biostatistics* (submitted)

Durbin, J. and Koopman, S.J. (2001). *Time Series Analysis by State Space Methods.* Oxford: Oxford University Press.

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

Laird, N.M. and Ware, J.H. (1982). Random-effects models for longitudinal data. *Biometrics*, **38**, 963–74.

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

Little, R.J. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data (seond edition).* New York: Wiley.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, **63**, 581–92.

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