A Short Course in Longitudinal Data Analysis

Peter Diggle, Gerwyn Green, Ines Sousa

(Department of Medicine, Lancaster University)

Lancaster, 19-20 June 2008
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 models
  15.00 BREAK
  15.30 Lab 2  The nlme package
  17.00 CLOSE
Day 2

  9.00  Lecture 3  Generalized linear models  
10.00  Lecture 4  Survival analysis  
11.00  BREAK  
11.30  Lab 3  The survival package  
13.00  Lunch  
13.30  Lecture 5  Joint modelling  
14.30  BREAK  
15.00  Lab 4  The joineR package  
17.00  CLOSE
Lecture 1. Introduction

- 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”

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Number of non-dropouts at week</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>haloperidol</td>
<td>85</td>
</tr>
<tr>
<td>placebo</td>
<td>88</td>
</tr>
<tr>
<td>risperidone</td>
<td>345</td>
</tr>
<tr>
<td>total</td>
<td>518</td>
</tr>
</tbody>
</table>
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

\[ Y_1, Y_2, ..., Y_t, ..., Y_n \quad Y_t \sim N(\mu, \sigma^2) \]

Classical result from elementary statistical theory:

\[ \bar{Y} \pm 2\sqrt{\frac{\sigma^2}{n}} \]

But if \( Y_t \) is a time series:

- \( \text{E}[\bar{Y}] = \mu \)

- \( \text{Var}\{\bar{Y}\} = (\sigma^2/n) \times \left\{ 1 + n^{-1} \sum_{u \neq t} \text{Corr}(Y_t, Y_u) \right\} \)
Correlation may or may not hurt you

\[ Y_{it} = \alpha + \beta(t - \bar{t}) + Z_{it} \quad i = 1, \ldots, m \quad t = 1, \ldots, n \]
Correlation may or may not hurt you

\[ Y_{it} = \alpha + \beta(t - \bar{t}) + Z_{it} \quad i = 1, \ldots, m \quad t = 1, \ldots, n \]
Correlation may or may not hurt you

\[ Y_{it} = \alpha + \beta(t - \bar{t}) + Z_{it} \quad i = 1, \ldots, m \quad t = 1, \ldots, n \]
Correlation may or may not hurt you

\[ Y_{it} = \alpha + \beta(t - \bar{t}) + Z_{it} \quad i = 1, \ldots, m \quad t = 1, \ldots, n \]

Parameter estimates and standard errors:

<table>
<thead>
<tr>
<th></th>
<th>ignoring correlation</th>
<th>recognising correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>estimate</td>
<td>standard error</td>
</tr>
<tr>
<td>( \alpha )</td>
<td>5.234</td>
<td>0.074</td>
</tr>
<tr>
<td>( \beta )</td>
<td>0.493</td>
<td>0.026</td>
</tr>
</tbody>
</table>
Balanced and unbalanced designs

\[ Y_{ij} = j^{th} \text{ measurement on } i^{th} \text{ subject} \]
\[ t_{ij} = \text{ time at which } Y_{ij} \text{ is measured} \]

- 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)

<table>
<thead>
<tr>
<th>Week</th>
<th>Mean</th>
<th>Variance</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>93.61</td>
<td>214.69</td>
<td>1.00</td>
</tr>
<tr>
<td>1</td>
<td>89.07</td>
<td>272.46</td>
<td>0.46</td>
</tr>
<tr>
<td>2</td>
<td>84.72</td>
<td>327.50</td>
<td>0.44</td>
</tr>
<tr>
<td>4</td>
<td>80.68</td>
<td>358.30</td>
<td>0.49</td>
</tr>
<tr>
<td>6</td>
<td>74.63</td>
<td>376.99</td>
<td>0.45</td>
</tr>
<tr>
<td>8</td>
<td>74.32</td>
<td>476.02</td>
<td>0.41</td>
</tr>
</tbody>
</table>
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

PANSS

time (weeks post-randomisation)
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

\[ r_{ij} = \text{residual from preliminary model for mean response} \]

• Define

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

• Estimate

\[ \hat{V}(u) = \text{average of all quantities } v_{iji\ell} \text{ such that } |t_{ij} - t_{i\ell}| \lesssim u \]

• Estimate of process variance

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

- 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
  \[ v_{ijil} \sim \{(s - 1)/s\}V(u_{ijil})\chi_1^2 \]

- \( \{s/(s - 1)\}\hat{V}(u) \) is unbiased for \( V(u) \)

- Expression available for covariance between any two quantities \( v_{ijil} \)

- Hence can compute variance of \( \hat{V}(u) \)

- Typically, \( \text{Var}\{\hat{V}(u)\} \) increases (sharply) as \( u \) increases
Where does the correlation come from?

- differences between subjects
- variation over time within subjects
- measurement error
Lab 1. Introduction to R

- invoking the R program
- the R mantra, $y = f(x)$
- scalar, vector and matrix arithmetic
- reading data from files
- graphics
- simple summary statistics
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. Linear models

- 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

\[ E(Y_{ij}) = x_{ij1}\beta_1 + \ldots + x_{ijp}\beta_p \]
\[ Y_i = X_i\beta + \epsilon_i \]
\[ Y = X\beta + \epsilon \]

• 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 (differences 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

\[ Y_{ij} - \mu_{ij} = U_i + Z_{ij} \]

\[ U_i \sim N(0, \nu^2) \]
\[ Z_{ij} \sim N(0, \tau^2) \]

Implies that \( \text{Corr}(Y_{ij}, Y_{ik}) = \frac{\nu^2}{\nu^2 + \tau^2} \), for all \( j \neq k \)
Random intercept and slope

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

\[(U_i, W_i) \sim \text{BVN}(0, \Sigma)\]

\[Z_{ij} \sim \text{N}(0, \tau^2)\]

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

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

\[ Y_{i1} - \mu_{i1} \sim \mathcal{N}\{0, \tau^2/(1 - \alpha^2)\} \]

\[ Z_{ij} \sim \mathcal{N}(0, \tau^2), \quad j = 2, 3, \ldots \]

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

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

\[ \text{Corr}\{W(t), W(t-u)\} = \rho(u) \]

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

\[ d_{ij} = \text{vector of explanatory variables for random effects} \]

\[ W_i(t) = \text{continuous-time Gaussian process} \]
(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

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

\[ V(u) = \tau^2 + \sigma^2 \{1 - \rho(u)\} \quad \text{Var}(Y_{ij}) = \nu^2 + \sigma^2 + \tau^2 \]
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:

• \( y = X\beta + \epsilon : V ar(\epsilon) = V \)

• \( \hat{\beta} = (X'X)^{-1}X'y \equiv Dy \)

• \( V ar(\hat{\beta}) = D\hat{V}D' \sim D\hat{V}D' \),
  \( \hat{V} = \) sample covariance matrix of OLS residuals

\[ \hat{\beta} \sim \text{MVN}(\beta, D\hat{V}D') \]
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 $\beta$ 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 $W$,  
- $\text{Var}(\tilde{\beta}_W) = \{(X'WX)^{-1}X'W\}V\{WX(X'WX)^{-1}\} = \Sigma$

Inference

$$\tilde{\beta}_W \sim N(\beta, \Sigma)$$
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 \((V_0 \text{ known})\)

Log-likelihood for observed data \(y\) is

\[
L(\beta, \sigma^2, V_0) = -0.5\{nm \log \sigma^2 + m \log |V_0| \\
+ \sigma^{-2}(y - X\beta)'(I \otimes V_0)^{-1}(y - X\beta)\}
\] (1)

Given \(V_0\), estimator for \(\beta\) is

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

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

Explicit estimator for \(\sigma^2\) also available as

\[
\hat{\sigma}^2(V_0) = \frac{RSS(V_0)}{(nm)}
\] (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

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

(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 $\sigma^2$ 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 $\sigma^2$ and $V_0$

• but
  
  – model for mean response then has $ng$ parameters
  
  – if $ng$ is large, maximum likelihood estimates of $\sigma^2$ 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,
  \[ Y \sim \text{MVN}\{X\beta, \sigma^2(I \otimes V_0)\}. \]

- transform data \( Y \) to \( Y^* = Ay \), matrix \( A \) chosen to make distribution of \( Y^* \) independent of \( \beta \).

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 $\beta$.

- estimate $\sigma^2$ and $V_0$ by maximising likelihood based on the transformed data $Y^*$. 
\[ \hat{\beta}(V_0) = (X'(I \otimes V_0)^{-1}X)^{-1}X'(I \otimes V_0)^{-1}Y \]

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

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

\[ = \mathcal{L}(V_0) - 0.5 \log |X'(I \otimes V_0)^{-1}X| \]

Note that:

- different choices for $A$ correspond to different rotations of coordinate axes within the residual space

- hence, REML estimates do not depend on $A$
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):** $P(\text{missing})$ depends neither on observed nor unobserved measurements

- **MAR (at random):** $P(\text{missing})$ depends on observed measurements, but not on unobserved measurements

- **MNAR (not at random):** conditional on observed measurements, $P(\text{missing})$ depends on unobserved measurements.
Example: Longitudinal clinical trial

- **completely at random**: patient leaves 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

\[ [Y, M] = [Y][M|Y] \]

• Pattern mixture factorisation

\[ [Y, M] = [M][Y|M] \]

• Random effects

\[ [Y, M] = \int [Y|U][M|U][U]dU \]
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
Longitudinal data with dropouts

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:

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

No \textit{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. \( P(D = d|\text{history}) = p_d(H_d, y^*_d; \phi). \)

- \( \beta \) parameterises mean response profile for \( y^* \)
- \( \alpha \) parameterises covariance structure of \( y^* \)
- \( \phi \) 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. \[
P(Y_k = 0|H_{k-1}, Y_{k-1} \neq 0) = \int p_k(H_k, y; \phi)f_k^*(y|H_k; \beta, \alpha)dy
\]

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 $Y = (Y_1, \ldots, Y_{d-1}, 0, \ldots, 0)$:

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

where $f^*_{d-1}(y)$ denotes joint pdf of $(Y^*_1, \ldots, Y^*_{d-1})$. 
Now consider a set of data with $m$ subjects.

- $\beta$ and $\alpha$ parameterise measurement process $y^*$
- $\phi$ parameterises dropout process

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

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

$$L_1(\beta, \alpha) = \sum_{i=1}^{m} \log\{f_{d_i-1}^*(y_i)\}$$

$$L_2(\phi) = \sum_{i=1}^{m} \sum_{k=1}^{d_i-1} \log\{1-p_k(H_{ik}, y_{ik})\}$$

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

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

If \( L_3(\cdot) \) only depends on \( \phi \), 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
\[ L_3(\beta, \alpha, \phi) = \sum_{i:d_i \leq n} \log\{P(Y_{id_i} = 0|H_{id_i} Y_{id_{i-1}} \neq 0)\}. \]

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

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

\[= p_k(H_k; \phi),\]

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 $\beta$ and $\alpha$ 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 $\rho$ 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.
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.
Lab 2. The \texttt{nlme} package

- linear regression modelling using ordinary least squares, the \texttt{lm()} function
- linear mixed effects modelling, the \texttt{lme()} function
- specifying longitudinal covariance structures
- analysis of data from a clinical trial
fit1=lm(CD4~time)
summary(fit1)

library(nlme)
?lme
fit2=lme(CD4~time,random=~1|id)
summary(fit2)
Lecture 3. Generalized linear models

- 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 $Y_i : i = 1, \ldots, n$.

1. $E(Y_i) = \mu_i : h(\mu_i) = x_i'\beta$, where $h(\cdot)$ is known link function, $x$ 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 $f(y; \mu_i, \phi)$
Two examples of classical GLM’s

Example 1: simple linear regression

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

- \( x_i = (1, d_i)' \)
- \( h(\mu) = \mu \)
- \( v(\mu) = 1, \phi = \sigma^2 \)
- \( f(y; \mu_i, \phi) = N(\mu_i, \phi) \)
Example 2: Bernoulli logistic model (binary response)

\[ P(Y_i = 1) = \frac{\exp(\beta_1 + \beta_2 d_i)}{1 + \exp(\beta_1 + \beta_2 d_i)} \]

- \( x_i = (1, d_i)' \)
- \( h(\mu) = \log\{\mu/(1 - \mu)\} \)
- \( v(\mu) = \mu(1 - \mu), \phi = 1 \)
- \( f(y; \mu_i) = \text{Bernoulli} \)
Three GLM constructions for longitudinal data

- random effects models
- transition models
- marginal models
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

- $E(Y_i|U) = \mu_i : h(\mu_i) = x_i'\beta + U'\alpha$
- $\text{Var}(Y_i|U) = \phi v(\mu_i)$
- $(Y_1, \ldots, Y_n)$ are mutually independent conditional on $U$.

Likelihood inference requires evaluation of

$$f(y) = \int \prod_{i=1}^{n} f(y_i|U)g(U)dU$$
Transition GLM

Conditional distribution of each $Y_i$ modelled directly in terms of preceding $Y_1, Y_2, \ldots, Y_{i-1}$.

- $E(Y_i|\text{history}) = \mu_i$

- $h(\mu_i) = x_i'\beta + Y_{(i)}'\alpha$, where $Y_{(i)} = (Y_1, \ldots, Y_{i-1})$

- $\text{Var}(Y_i|\text{history}) = \phi v(\mu_i)$

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

$$f_k(y_k|y_1, \ldots, y_{k-1}) = f_k(y_k|y_{k-1})$$

and condition on $y_1$ as model does not directly specify $f_1(y_1)$. 
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 $\beta$ uses method of generalized estimating equations
(the clever ostrich revisited)
Estimating equations

Estimating equations for $\beta$ 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 $\frac{\partial \mu_i}{\partial \beta_j}$
- $V_i$ is an $n \times n$ diagonal matrix with non-zero elements proportional to $\text{Var}(Y_i)$
- $Y$ and $\mu$ are $n$-element vectors with elements $Y_i$ and $\mu_i$
Generalized estimating equations (GEE)

In longitudinal setting:

- single $n$-element vectors $Y$ and $\mu$ replaced by $n_i$-element vectors $Y_i$ and $\mu_i$, 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 $\hat{\beta}$

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

(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

\[ \sqrt{(m)}(\hat{\beta} - \beta) \sim MVN(0, I_0^{-1} I_1 I_0^{-1}) \]  \hspace{1cm} (6)

where

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

and

\[ I_1 = \sum_{i=1}^{m} \{D_{\mu_i\beta}\}'\{V_i(\alpha)\}^{-1}\text{Var}(Y_i)\{V_i(\alpha)\}^{-1}D_{\mu_i\beta} \]
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 $\mu_i$ 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, $\beta$ measures population-averaged effects of explanatory variables on mean response

• in transition or random effects modelling, $\beta$ 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*runif(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)
Lecture 4. Survival analysis

• lifetime distributions and hazard rates
• estimating a survival curve, censoring, Kaplan-Meir
• parametric models
• Cox proportional hazards models
• frailty
How long until the next bus?

- a Poisson process
• the waiting time distribution

\[ \lambda = \text{average number of buses per unit time} \]

• the paradox: however long you have been waiting, the probability distribution of the time you still have to wait is the same as when you arrived at the bus-stop
Resolving the paradox

\[ T = \text{time until the event of interest occurs} \]

- the probability density function, \( f(t) \)
  \[ f(t) \times h \approx P(\text{event occurs between } t \text{ and } t + h) \]

- the survivor function \( S(t) \)
  \[ S(t) = \int_t^\infty f(s)ds \]
  \[ S(t) = P(\text{event has not occurred by time } t) \]

- the hazard function \( \lambda(t) \)
  \[ \lambda(t) = \frac{f(t)}{S(t)} \]
  \[ f(t) \times h \approx P(\text{event occurs between } t \text{ and } t + h), \]
  conditional on its not having occurred before time \( t \)

If \( f(t) = \lambda \exp(-\lambda t) \) then \( \lambda(t) = \lambda \) (constant)
Survival data

Data: 1.5 0.5 3.8 2.1 4.6 0.7 0.2 2.8 7.3 5.9
Estimating the survivor function

Data: 1.5 0.5 3.8 2.1 4.6 0.7 0.2 2.8 7.3 5.9
Estimating the survivor function

Data: 1.5 0.5 3.8 2.1 4.6 0.7 0.2 2.8 7.3 5.9
Censoring

In most applications, the study ends before all subjects have experienced the event of interest
The Kaplan-Meir estimate

Data: 0.2 0.5 0.7+ 1.5 2.1 2.8 3.8 4.6+ 5.9 7.3+

<table>
<thead>
<tr>
<th>time</th>
<th>survivor curve</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>$\frac{10}{10}$ = 1.000</td>
</tr>
<tr>
<td>0.20</td>
<td>$\frac{10}{10} \times \frac{9}{10}$ = 0.900</td>
</tr>
<tr>
<td>0.50</td>
<td>$\frac{10}{10} \times \frac{9}{10} \times \frac{8}{9}$ = 0.800</td>
</tr>
<tr>
<td>0.70</td>
<td>$\frac{10}{10} \times \frac{9}{10} \times \frac{8}{9}$ = 0.800</td>
</tr>
<tr>
<td>1.50</td>
<td>$\frac{10}{10} \times \frac{9}{10} \times \frac{8}{9} \times \frac{6}{7}$ = 0.686</td>
</tr>
</tbody>
</table>
The Kaplan-Meir curve

Data: 0.2 0.5 0.7+ 1.5 2.1 2.8 3.8 4.6+ 5.9 7.3+
Parametric models

- gamma
- Weibull
- log-Gaussian
Parametric models: gamma distribution

$f(t)$ (left-hand panel) and $\lambda(t)$ (right-hand panel)
Adding explanatory variables

- Linear models assume that explanatory variables shift the distribution to the left or right:
in survival analysis, this often isn’t a natural assumption because the survival time distribution has to be anchored at zero.

two popular alternatives are:

- accelerated life
  \( S_k(t) = \) survivor function in treatment group \( k \)
  \( S_k(t) = S_0(\alpha_k t) : k = 1, 2, ..., m \)

- proportional hazards
  \( \lambda_k(t) = \) survivor function in treatment group \( k \)
  \( \lambda_k(t) = \lambda_0(t) \times \alpha_k \)
Cox proportional hazards models

Cox (1972) revolutionised survival analysis by introducing the following class of models:

\[ \lambda_i(t) = \lambda_0(t) \exp\{x_i'\beta\} \]

where:

- \( x_i \) is a vector of explanatory variables for \( i^{th} \) subject;
- \( \beta \) is a vector of regression parameters;
- \( \lambda_0(t) \) is an arbitrary baseline hazard function
Estimation for Cox proportional hazards

- subjects: $i = 1, \ldots, n$
- explanatory variables: $x_i : i = 1, \ldots, n$
- hazard functions: $\lambda_i(t) = \lambda_0(t) \exp(x_i' \beta)$
\[ \lambda_i(t) = \lambda_0(t) \exp(x'_i \beta) \]

Partial likelihood: condition on the ordered survival and censoring times, and calculate the probability of the observed ordering

- observed survival times \( t_1 < t_2 < \ldots < t_n \)
- censored survival times \( c_{n+1} < c_{n+2} < \ldots < c_{n+m} \)
- \( R(t) \) = risk set at time \( t \) (all subjects with \( t_i \geq t \) or \( c_i > t \))
- \( p_i = \lambda_i(t_i) / \sum_{j \in R(t_i)} \lambda_j(t_i) \)
- partial log-likelihood: \( PL(\beta) = \sum_{i=1}^{n} \log p_i \)

Partial log-likelihood does not depend on \( \lambda_0(t) \)
Modelling the PANSS dropout times

Fitted survivor curves for log-Gaussian (left-hand panel) and non-parametric (right-hand panel) hazards
Frailty

- Frailty describes the situation in which the variability in the survival times of different subjects is bigger than can be explained by the explanatory variables $x$

- One way to deal with this is to imagine a “missing explanatory variable” $U$ and extend the proportional hazards model to

$$\lambda_i(t) = \lambda_0(t) \exp(x_i'\beta + U_i)$$

- since we can’t observe the values of the $U_i$, we assume that they are a random sample from a frailty distribution
Frailty

- two popular choices for the frailty distribution are

1. $U \sim \mathcal{N}(0, \sigma^2)$
2. $\exp(U) \sim \Gamma(1, \kappa)$

More on frailty in Lecture 5 (joint modelling)
Lab 3. The survival package

- Kaplan-Meir plots
- adding confidence limits to a Kaplan-Meir plot
- fitting Cox proportional hazards models using the coxph function
set.seed(5968)
F=exp(1+0.5*rnorm(100))
q=1-(1:100)/100
plot(sort(F),q,xlab="t",ylab="S(t)")

library(survival)
alive=(runif(100)<0.5)
table(alive)
Kaplan.Meir=survfit(Surv(F,censored))
plot(Kaplan.Meir,conf.int=FALSE,xlab="t",ylab="S(t)")
lines(sort(F),q,col="red")
Lecture 5. Joint modelling

- what is it?
- why do it?
- random effects models
- transformation models
Joint modelling: what is it?

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

$[Y, F|x, \theta]$
Prothrombin index data

- Placebo-controlled RCT of prednisone for liver cirrhosis patients.
  Total $m = 488$ subjects.

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

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

  Andersen, Borgan, Gill and Keiding, 1993
Smoothed mean prothrombin

Prothrombin index vs. Time (years)

- **Treatment** (n=251)
- **Control** (n=237)
Kaplan–Meier

Treatment (n=251)
Control (n=237)
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:

<table>
<thead>
<tr>
<th>week</th>
<th>−1</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>6</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>missing</td>
<td>0</td>
<td>3</td>
<td>9</td>
<td>70</td>
<td>122</td>
<td>205</td>
<td>251</td>
</tr>
<tr>
<td>proportion</td>
<td>0.00</td>
<td>0.01</td>
<td>0.02</td>
<td>0.14</td>
<td>0.24</td>
<td>0.40</td>
<td>0.49</td>
</tr>
</tbody>
</table>

- 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, 2006
Heart surgery data
Mean log-LVMI response profiles
Heart surgery data
Survival curves adjusted for baseline covariates

Cox Proportional model, 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 frailty 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

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

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

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
- \( \eta_2(t) = X_2(t) + R_{2i}(t) = \) 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

Two choices for measurement process component:

\[ R_1(t) = U_1 + W_1(t) \]
\[ R_1(t) = U_1 + U_2t \]

And for hazard process component:

\[ R_2(t) = \gamma_1 R_1(t) \]
\[ R_2(t) = \gamma_1(U_1 + U_2t) + \gamma_2 U_2 \]
\[ = \gamma_1 R_1(t) + \gamma_2 U_2 \]
Schizophrenia trial data
Mean response (random effects model)
Schizophrenia trial data
Empirical and fitted variograms
A simple transformation model

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

• 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 $[Y, S] = \text{MVN}(\mu, \Sigma)$
- Use factorisation $[Y, S] = [Y][S|Y]$
- $\mu = (\mu_Y, \mu_S)$
- Standard result for $[S|Y]$
  
  - $S|Y \sim \text{N}(\mu_{S|Y}, \sigma^2_{S|Y})$
  - $\mu_{S|Y} = \mu_S + g'(\phi)V(\theta)^{-1}(Y - \mu_Y)$
  - $\sigma^2_{S|Y} = \nu^2 - g'(\phi)V(\theta)^{-1}g(\phi)$
Missing values and censoring

- **uncensored** $S_i$:
  \[ [Y_i] \times [S_i | Y_i] \]

- **right-censored** $S_i > t_{ij}$
  \[ [Y_i] \times [1 - \Phi\{ (t_{ij} - \mu_{S|Y_i}) / \sigma_{S|Y} \}] \]

- **interval-censored** $t_{ij} < S_i < t_{i,j+1}$
  \[ [Y_i] \times [\Phi\{ (t_{i,j+1} - \mu_{S|Y_i}) / \sigma_{S|Y} \} - \Phi\{ (t_{ij} - \mu_{S|Y_i}) / \sigma_{S|Y} \}] \]

- **missing** $Y_{ij}$
  - reduce dimensionality of $Y_i$ 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:
  - \( \text{Var}(Y) = V(\theta) \)
  - \( \text{Var}(S) = \nu^2 \)
  - \( g(\phi) = \text{Cov}(Y, S) \)

• Standard choices for \( V(\theta) \) include:
  - Random intercept and slope (Laird and Ware, 1982)
    \[ Y_{ij} - \mu_{ij} = A_i + B_i t_{ij} + Z_{ij} : j = 1, \ldots, n_i; i = 1, \ldots, m \]
  - Three components of variation (Diggle, 1988)
    \[ Y_{ij} - \mu_{ij} = A_i + W_i(t_{ij}) + Z_{ij} \]
  - Compound symmetry
    \[ Y_{ij} - \mu_{ij} = A_i + Z_{ij} \]
• 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 \( N(0, 1) \) samples.
Re-analysis of schizophrenia trial data

Dropout is not completely at random

subject about to drop-out

subject not about to drop-out
Re-analysis of schizophrenia trial data
Model specification

- measurements, $Y$: random intercept and slope
  \[ Y_{ij} - \mu_{ij} = A_i + B_i t_{ij} + Z_{ij} : j = 1, \ldots, n_i; i = 1, \ldots, m \]

- dropout time, $F$
  \[ S = \log F \sim N(\mu_S, \nu^2) \]

- cross-covariances
  \[ \text{Cov}(Y_j, S) = \phi_j : j = 1, \ldots, 6 \]
Re-analysis of schizophrenia trial data
Goodness-of-fit: subjects about to drop out
Re-analysis of schizophrenia trial data
Goodness-of-fit: not about to drop out
Re-analysis of schizophrenia trial data
Goodness-of-fit: mean response profiles
Re-analysis of schizophrenia trial data
Fitted mean response profiles

haloperidol

placebo

risperidone
Lab 4. The joiner package

- reading data-files
- exploratory analysis
- random effects models
- transformation models
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 construction
- 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
References


