NTNU
Norwegian University of
Science and Technology

KLMED8008 Analysis of repeated measurements: Longitudinal data, Ch 5 and 6

Turid Follestad

Multilevel nested data

So far: Two-level nested data
Examples:

- patients in hospitals
- siblings in families
- twins in twin-pairs

Three-level nested data:

- patients in hospitals in countries
- siblings in families in districts
- occasions in twins in twin-pairs

Two-level nested data

- The data are nested in the sense that a lower level unit can only belong to one higher level unit (cluster).
- There is in general no natural ordering of the units within each cluster

Random intercept models

$$y_{i,j} = (\beta_1 + \zeta_j) + \beta_2 x_{i,j} + \epsilon_{i,j}$$
$$\zeta_j \sim N(0, \psi)$$
$$\epsilon_{i,j} \sim N(0, \theta)$$
$$j = \text{cluster}, \ i = \text{unit within cluster}$$

- Random intercepts $\zeta_j; j = 1, \ldots, J$ are independent
- Residuals $\epsilon_{i,j}; j = 1, \ldots, J, i = 1, \ldots, n_j$ are independent, and independent of the random intercepts

The joint distribution of the observables $\{y_{i,j}\}$,

$$f(\{y_{i,j}\}; \beta_1, \beta_2, \psi, \theta),$$

is fully specified, given these assumptions and the model parameters.

Random coefficient models

$$y_{i,j} = (\beta_1 + \zeta_{1,j}) + (\beta_2 + \zeta_{2,j})x_{i,j} + \epsilon_{i,j}$$

The pair $\zeta_{\mathbf{j}} = (\zeta_{1,j}, \zeta_{2,j})$ is assumed bivariate normal with mean zero and covariance matrix

$$\Psi = \begin{pmatrix} \psi_{1,1} & \psi_{1,2} \\ & \psi_{2,2} \end{pmatrix}, \ \psi_{1,2} = \sqrt{\psi_{1,1}}\sqrt{\psi_{2,2}}\,\rho_{1,2}, \ \rho_{1,2} = \text{correlation}$$

$$\epsilon_{i,j} \sim N(0, \theta)$$

— The pairs $\zeta_{\mathbf{j}} = (\zeta_{1,j}, \zeta_{2,j})$ are independent across clusters
— The residuals $\epsilon_{i,j}$ are independent within and between clusters

The joint distribution of the observables $\{y_{i,j}\}$,

$$f(\{y_{i,j}\}; \beta_1, \beta_2, \psi_{1,1}, \psi_{2,2}, \rho, \theta),$$

is fully specified, given these assumptions and the parameters.

Random intercept and random coefficient models: Illustration

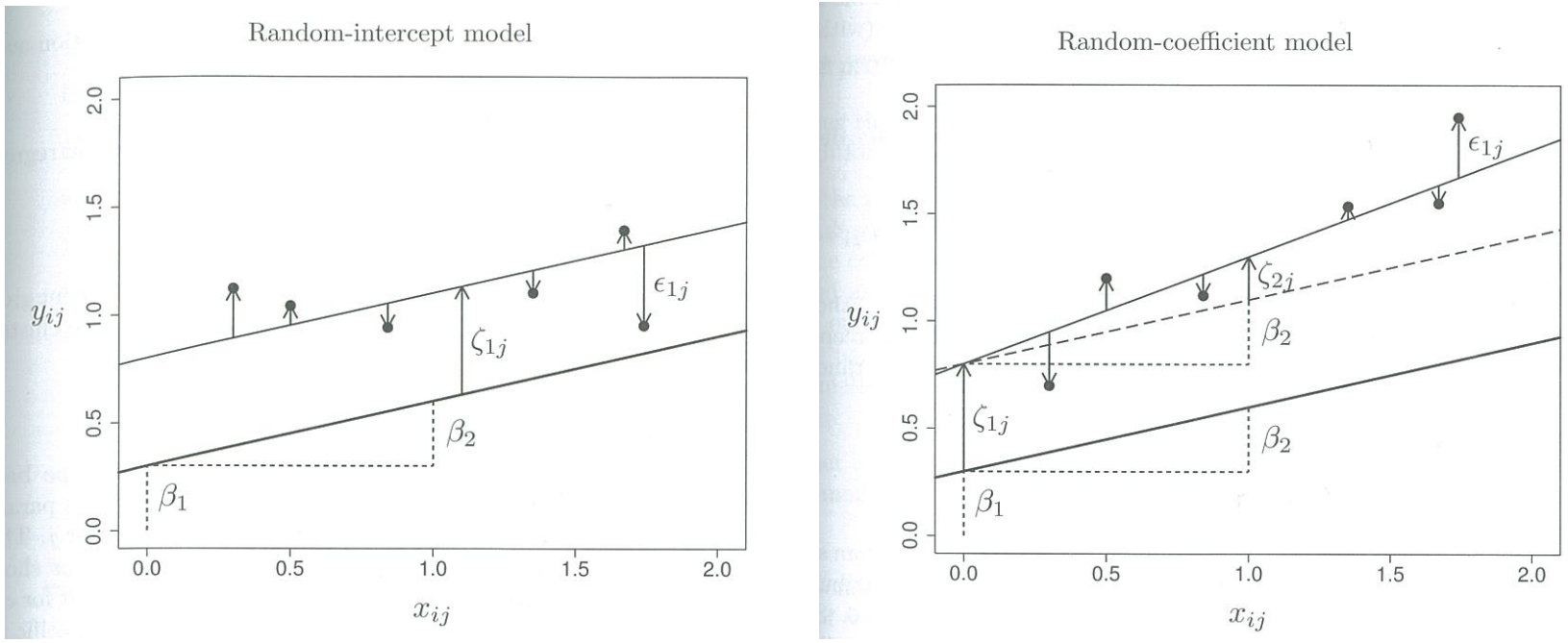


Fig 4.5 in Rabe-Hesketh and Skrondal.

Longitudinal data

- Subjects repeatedly measured over time
- Also called *panel data*, *repeated measures*, *cross-sectional time series data*
- Examples:
  - Outcomes of a randomized clinical trial (RCT) comparing two treatments, recorded over time
  - Students and their standardized test scores in six successive years.
  - Hourly wages and explanatory variables (such as years of education) recorded over years (example in Ch 5 in textbook)

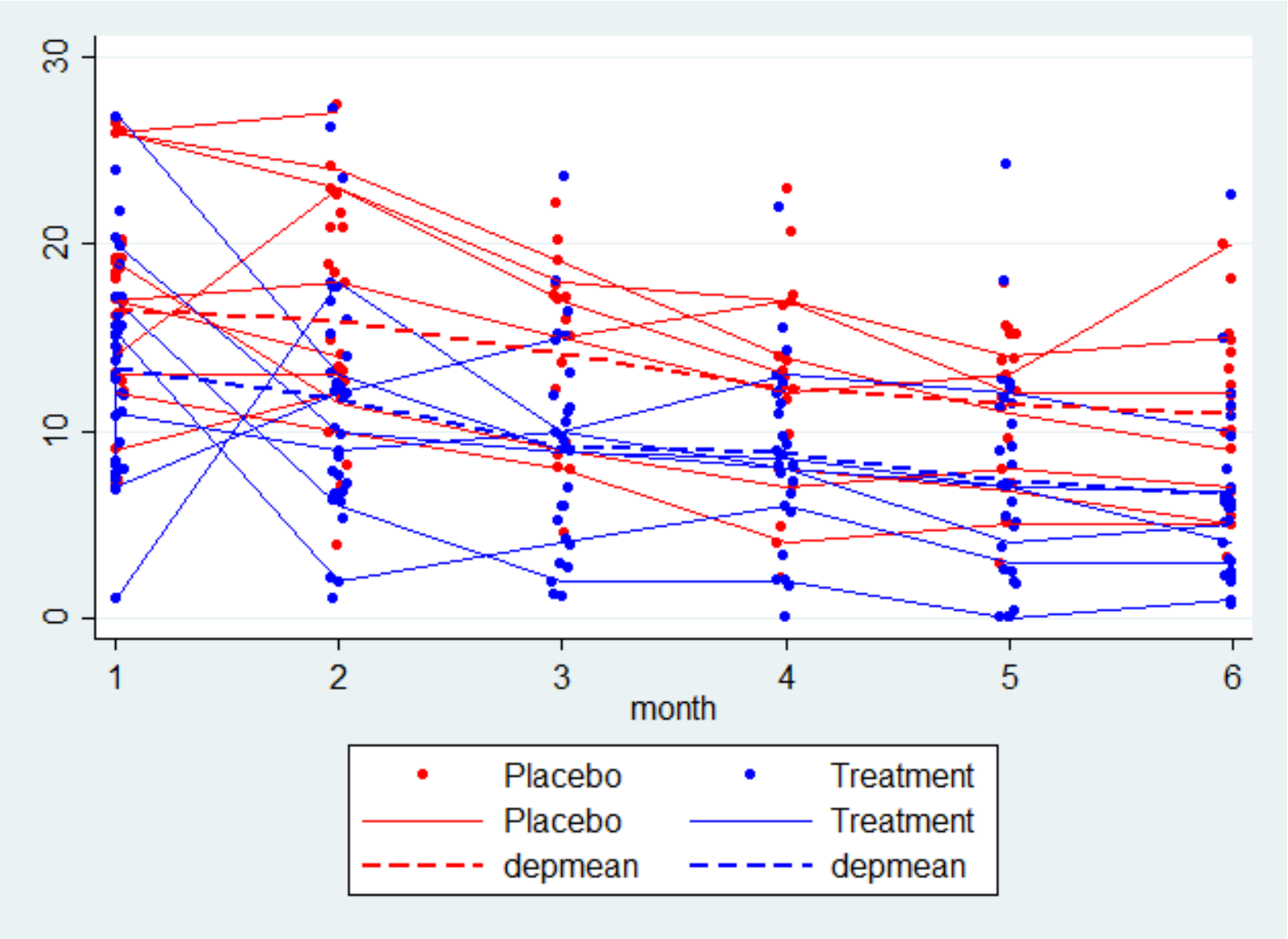
Longitudinal study designs

$t_{1,1}$	$t_{1,2}$	$t_{1,3}$	$t_{1,4}$	$t_{1,5}$
$t_{2,1}$	$t_{2,2}$	$t_{2,3}$	$t_{2,4}$	$t_{2,5}$
∘	∘	∘	∘	∘
∘	∘	∘	∘	∘
∘	∘	∘	∘	∘
$t_{n,1}$	$t_{n,2}$	$t_{n,3}$	$t_{n,4}$	$t_{n,5}$

$$t_{i,j} = t_{\text{subject } i, \text{time } j}$$

- Panel studies: All individuals measured at the same occasions ($t_{i,j} = t_j$, balanced data)
- Cohort studies: A group of subjects measured at, in general, subject-specific occasions
- RCT: The individuals are randomly assigned to treatments, balanced data if possible

Example (exercise 6.2)



Example: Missing data

```
. xtset subj month
       panel variable:  subj (strongly balanced)
        time variable:  month, 1 to 6
                delta:  1 unit

. xtdescribe if dep<.

    subj:  1, 2, ...,  61                                        n =         61
   month:  1, 2, ...,  6                                         T =          6
           Delta(month) = 1 unit
           Span(month)  = 6 periods
           (subj*month uniquely identifies each observation)

Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                         1       1       3         6         6       6       6

     Freq.  Percent    Cum. |  Pattern
 ---------------------------+-----------
        45    73.77   73.77 |  111111
         8    13.11   86.89 |  1.....
         7    11.48   98.36 |  11....
         1     1.64  100.00 |  111...
 ---------------------------+-----------
        61   100.00         |  XXXXXX

.
```

Missing data

Missing data in longitudinal studies can be

- Drop-outs
- Intermittent missing values

Types of missing data mechanisms:

- Missing at random (MAR)
- Missing not at random (MNAR)

LMM gives consistent estimates of the model parameters as long as data are MAR.

Longitudinal data: Correlations

— Two-level, clustered data:
  - No ordering of the units within clusters, units are *exchangeable*
  - Residuals often assumed independent

— Longitudinal data:
  - Clusters = subjects, units = occasions, units ordered within clusters.
  - Often within-subject correlations (auto-correlations, serial correlations) in addition to those induced by random intercept and slope.

In the special case that the repeated measurements given covariates are exchangeable (order does not matter), the data can also be viewed as two-level data with measurements nested in subjects.

Why special models for longitudinal data?

Why longitudinal data?

- Causal inference (not possible with cross-sectional data)
- Each subject can act as his/hers own control in within-subject comparisons.
- Growth-curve models: A special case of random coefficient models with covariates being functions (often polynomials or piecewise linear functions) of time.

Why special models?

- Wiithin-subject residual correlations
- Consistent estimates in case of missing data

Approaches for modeling longitudinal data

- Random effects models (Ch 5)
- Marginal models (Ch 6)
- Other methods (predominant in ecomonics): Fixed-effects models and dynamical or lagged-response models

Random effects models for longitudinal data

Random intercept model, time as factor:

$$\begin{aligned} y_{i,j} &= \alpha_i + \beta x_{i,j} + \zeta_j + \epsilon_{i,j} \\ \alpha_i &= \text{effect of time} \end{aligned}$$

Random intercept model, time as covariate (linear predictor):

$$\begin{aligned} y_{i,j} &= \beta_1 + \alpha t_i + \beta_2 x_{i,j} + \zeta_j + \epsilon_{i,j} \\ \alpha &= \text{fixed slope for time} \end{aligned}$$

$$\zeta_j \sim N(0, \psi)$$

— Clustered data: $\{\epsilon_{i,j}\}$ assumed to be independent within and between clusters, $\epsilon_{i,j} \sim N(0, \theta)$

— Longitudinal data: $\epsilon_j = (\epsilon_{1,j}, \epsilon_{2,j}, \ldots, \epsilon_{n_j})$ multivariate normal with mean zero and covariance matrix Σ.

Joint distribution of observables $\{y_{i,j}\}$: $f(\{y_{i,j}\}; \beta, \{\alpha_i\}, \psi, \Sigma)$

Example: RCT

- A randomized clinical trial with n_t occasions, including baseline
- Two factors: Time and treatment (2 levels) with interaction

Random intercept model:

$$y_{time,id} = \alpha_{time,treatment(id)} + \zeta_{id} + \epsilon_{time,id}$$

By randomization:

$$\alpha_{\text{baseline},1} = \alpha_{\text{baseline},2}$$

- $\zeta_{id} \sim N(0, \psi)$
- Covariance matrix Σ of $\epsilon_{\mathbf{id}} = (\epsilon_{1,id}, \epsilon_{2,id}, \ldots, \epsilon_{n_t,id})$.

Multi-level vs marginal models

Example: Random intercept model

$$y_{time,id} = \alpha_{time,treatment(id)} + \zeta_{id} + \epsilon_{time,id}$$

Multi-level model: Specify

— $\zeta_{id} \sim N(0, \psi)$

— Covariance matrix Σ of $\epsilon_{\mathbf{id}} = (\epsilon_{1,id}, \epsilon_{2,id}, \ldots, \epsilon_{n_t,id})$.

Marginal model: Specify

— Covariance matrix V of $\mathbf{y_{id}} = (y_{1,id}, y_{2,id}, \ldots, y_{n_t,id})$.

- V equals covariance matrix for the *total residuals* $\xi_{time,id}$
- For random intercept model: $\xi_{time,id} = \zeta_{id} + \epsilon_{time,id}$

Multi-level vs marginal models

Multi-level model: Focus is on

- Population-averaged relationships between response and covariates: Fixed part of the model
- Subject-specific relationships: Combination of fixed and subject-specific random effects
- Variability of subject-specific effects around the population average

Marginal model: Focus is on

- Population-averaged relationships between response and covariates: Fixed part of the model
- Marginal residual covariance matrix, i.e. covariance matrix of the response $y_{i,j}$, or equivalently the total residuals

Common specifications of the covariance matrix V (marginal model), or Σ (multi-level model):

— Identity covariance structure:

$$V = \sigma^2 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

— Unstructured covariance:

$$V = \begin{pmatrix} a & b & c \\ b & d & e \\ c & e & f \end{pmatrix}$$

— Autoregressive covariance structure of order 1, AR(1):

$$V = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{pmatrix}$$

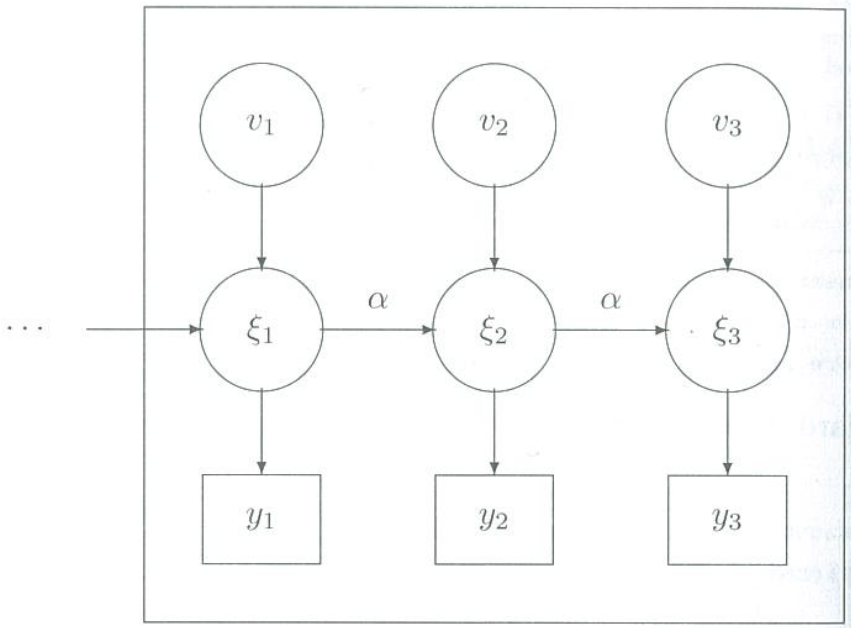


Figure 6.4: Path diagram of AR(1) process

$y_{i,j} = \text{fixed part} + \xi_{i,j}$
$\xi_{i,j} = \alpha\, \xi_{i-1,j} + v_{i,j}$
$(\alpha = \rho)$

— Meaningful only if time points are regularly spaced.
— Commonly used model: Random intercept + AR(1) model for within-subject residuals $\epsilon_{i,j}$ (AR(1)-model for Σ)

— Compound symmetry, or exchangeable:

$$V = \sigma^2 \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}$$

> Marginal covariance structure from random intercept model has compound symmetry:

$$\begin{aligned} V &= \psi_{1,1} \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} + \theta \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} \psi_{1,1} + \theta & \psi_{1,1} & \psi_{1,1} \\ \psi_{1,1} & \psi_{1,1} + \theta & \psi_{1,1} \\ \psi_{1,1} & \psi_{1,1} & \psi_{1,1} + \theta \end{pmatrix} \end{aligned}$$

— Exponential covariance structure:

$$V = \sigma^2 \begin{pmatrix} 1 & \rho^{t_2-t_1} & \rho^{t_3-t_1} \\ \rho^{t_2-t_1} & 1 & \rho^{t_3-t_2} \\ \rho^{t_3-t_1} & \rho^{t_3-t_2} & 1 \end{pmatrix}$$

— Toeplitz covariance structure:

$$V = \begin{pmatrix} a & b & c \\ b & a & b \\ c & b & a \end{pmatrix}$$