Longitudinal Data Analysis

methods@manchester summer school

Day 2 | morning session

Thiago R. Oliveira

Lecturer in Quantitative Criminology, University of Manchester

30/06—04/07

Today

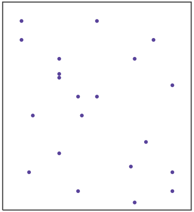
Growth curve models: a multilevel approach

- Motivation: Why multilevel models?
- Within and between variation
- Pooled, fixed effects, and random effects models
- Multilevel models for longitudinal data
- Growth curve models

Multilevel models

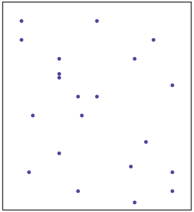
Idea of multilevel data

⇝ Assume a pool of structured data



Idea of multilevel data

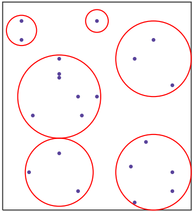
- ⇝ Assume a pool of structured data



- ⇝ Each dot represents a unit i

Idea of multilevel data

⇝ Assume a pool of structured data



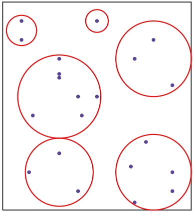
⇝ Each dot represents a unit i

⇝ Each circle represents a group j

⇝ Observations are often clustered (e.g., students within schools, residents within neighbourhoods).

Idea of multilevel data

- Assume a pool of structured data



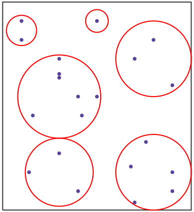
- Each dot represents a unit i
- Each circle represents a group j

· Pooled approach

- Observations are often clustered (e.g., students within schools, residents within neighbourhoods).
- Standard regression assumes independence across all observations — often violated in hierarchical or longitudinal data.

Idea of multilevel data

- Assume a pool of structured data



- Each dot represents a unit i
- Each circle represents a group j

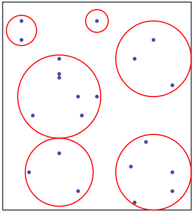
· Pooled approach

· Between approach

- Observations are often clustered (e.g., students within schools, residents within neighbourhoods).
- Standard regression assumes independence across all observations — often violated in hierarchical or longitudinal data.

Idea of multilevel data

⇝ Assume a pool of structured data



⇝ Each dot represents a unit i

⇝ Each circle represents a group j

· Pooled approach

· Between approach

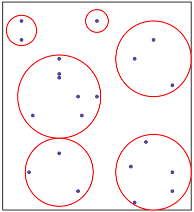
· Fixed Effects

⇝ Observations are often clustered (e.g., students within schools, residents within neighbourhoods).

⇝ Standard regression assumes independence across all observations — often violated in hierarchical or longitudinal data.

Idea of multilevel data

⇝ Assume a pool of structured data



⇝ Each dot represents a unit i

⇝ Each circle represents a group j

· Pooled approach

· Between approach

· Fixed Effects

· Random Effects

⇝ Observations are often clustered (e.g., students within schools, residents within neighbourhoods).

⇝ Standard regression assumes independence across all observations — often violated in hierarchical or longitudinal data.

⇝ Multilevel models allow us to model this structure explicitly.

Within vs Between Variation

- **Within variation**: differences within the same group
  - e.g., how residents of a neighbourhood are exposed to urban violence
- **Between variation**: differences between groups
  - e.g., how neighbourhoods differ from one another
- Multilevel models allow us to separate and model both sources of variation.

Modelling options

- **Pooled model:** Ignores grouping structure
- **Aggregate analysis:** Between variation only
- **Fixed effects model:** Within variation only (*very powerful for causal inference!*)
- **Random effects model:** Models heterogeneity as random variables

Within vs Between Variation

- **Within variation**: differences within the same group
  - e.g., how residents of a neighbourhood are exposed to urban violence
- **Between variation**: differences between groups
  - e.g., how neighbourhoods differ from one another
- Multilevel models allow us to separate and model both sources of variation.

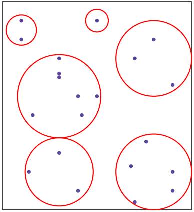
Modelling options

- **Pooled model:** Ignores grouping structure
- **Aggregate analysis:** Between variation only
- **Fixed effects model:** Within variation only (*very powerful for causal inference!*)
- **Random effects model:** Models heterogeneity as random variables

Multilevel models for panel data

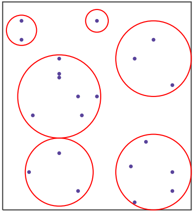
Longitudinal data as multilevel data

- Assume a pool of structured data



Longitudinal data as multilevel data

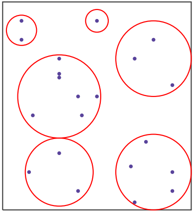
⇝ Assume a pool of structured data



⇝ Each circle represents an **individual** i

Longitudinal data as multilevel data

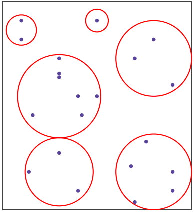
- ⇝ Assume a pool of structured data



- ⇝ Each dot represents an **observation** t
- ⇝ Each circle represents an **individual** i

Longitudinal data as multilevel data

⇝ Assume a pool of structured data



⇝ Each dot represents an **observation** t

⇝ Each circle represents an **individual** i

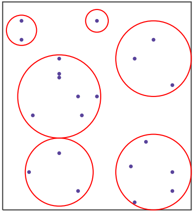
⇒ **Level 1**: Observations

⇒ **Level 2**: Individuals

⇝ *Observations are nested within individuals*

Longitudinal data as multilevel data

⇝ Assume a pool of structured data



⇝ Each dot represents an **observation** t

⇝ Each circle represents an **individual** i

⇒ **Level 1**: Observations

⇒ **Level 2**: Individuals

⇝ *Observations are nested within individuals*

⇝ **Within variation**: *change over time*

· e.g., how each student's test scores evolve over time

⇝ **Between variation**: *differences between individuals*

· e.g., how students differ from one another

Growth Curve Models: basic setup

Growth Curve Models

- a.k.a. latent growth/trajectory models
- Goal is to estimate developmental trajectories over time
- Basic setup
  - Dependent variable: An outcome of interest
  - Independent variable: *time*
  - Also assuming there is an underlying distribution driving individual trajectories: **growth parameters**
- Equivalent growth curve models can be viewed as multilevel models or as structural equation models (*more on that later)*
  - For a multilevel approach, we set the data in a *long* format with nT observations and assume that observations are nested within individuals

Growth Curve Models

Let's start with an example

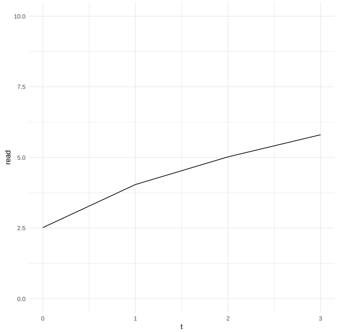
- Data from children of female respondents to the US National Longitudinal Survey of Youth
- Reading scores for 221 children on four occasions (two years apart from 1986 to 1992)

childid (i)	year (t)	male (x_i)	read (y_{it})
1	1	1	2.1
1	2	1	2.9
1	3	1	4.5
1	4	1	4.5
2	1	0	2.3
2	2	0	4.5
2	3	0	4.2
2	4	0	4.6

Table: Reading scores (y_{it}) over time for two children

Growth Curve Models

	Year			
	1	2	3	4
Mean	2.52	4.04	5.02	5.80

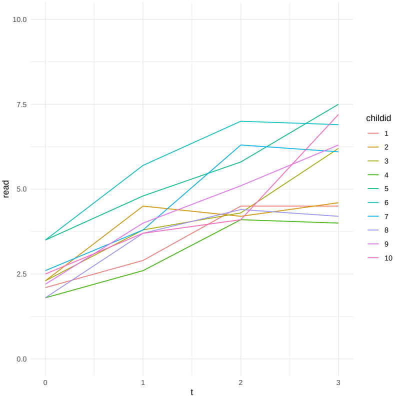


Overall, reading score increases by year

But there probably is a large amount of variation between children

Growth Curve Models

Let's randomly pick **10 children** and analyse their individual trajectories over time



⇒ Individual variation in level (intercept) and rate of change (slope)

⇒ In growth curve modelling, we aim to fit a curve that captures the observed variation *within* and *between* children as closely as possible

Growth Curve Models

Types of questions we could make

- What is the nature of reading development with age? Linear or nonlinear?
- How much do children vary in their initial reading score and in their rate of development?
- Does the initial score and rate of change depend on child/family characteristics?

Growth Curve Models: linear and non-linear trajectories

Basic linear growth curve model

The basic setup is regressing an outcome y_{it} on time t_{it} (let's ignore covariates for now)

$$y_{it} = \alpha_i + \beta_1 \cdot t_{it} + \epsilon_{it}$$

Note the subscript in α_i. **The intercept is allowed to vary by individuals i.**

$\Rightarrow$ α_i is a random effect! It implies that each individual is allowed to have their own initial score in the trajectory

- $\rightsquigarrow$ Known as a *Random Intercepts Model*
- $\rightsquigarrow$ α_i can be further decomposed in $\alpha_i = \bar{\alpha} + \mu_i$
  - $\bar{\alpha}$ is the *average* intercept, μ_i is the subject-specific residual

Basic linear growth curve model

The basic setup is regressing an outcome y_{it} on time t_{it} (let's ignore covariates for now)

$$y_{it} = \alpha_i + \beta_1 \cdot t_{it} + \epsilon_{it}$$

Note the subscript in α_i. **The intercept is allowed to vary by individuals i.**

$\Rightarrow$ α_i is a random effect! It implies that each individual is allowed to have their own initial score in the trajectory

- $\rightsquigarrow$ Known as a *Random Intercepts Model*
- $\rightsquigarrow$ α_i can be further decomposed in $\alpha_i = \bar{\alpha} + \mu_i$
  - $\bar{\alpha}$ is the *average* intercept, μ_i is the subject-specific residual

Basic linear growth curve model

The basic setup is regressing an outcome y_{it} on time t_{it} (let's ignore covariates for now)

$$y_{it} = \alpha_i + \beta_1 \cdot t_{it} + \epsilon_{it}$$

Note the subscript in α_i. **The intercept is allowed to vary by individuals i.**

$\Rightarrow$ α_i is a random effect! It implies that each individual is allowed to have their own initial score in the trajectory

- $\rightsquigarrow$ Known as a *Random Intercepts Model*
- $\rightsquigarrow$ α_i can be further decomposed in $\alpha_i = \bar{\alpha} + \mu_i$
  - $\bar{\alpha}$ is the *average* intercept, μ_i is the subject-specific residual

Basic linear growth curve model

Interpretation of terms

$$y_{it} = \alpha_i + \beta_1 \cdot t_{it} + \epsilon_{it}$$

- $y_{it} = \bar{\alpha} + \beta_1 \cdot t_{it}$ is the overall trajectory across all individuals
- $\bar{\alpha}$ is the expected value of y at $t_{it} = 0$ across all individuals
- β_1 is the effect of a one-unit increase in time on y
- μ_i is the departure of y for individual i from the overall trajectory

Basic linear growth curve model

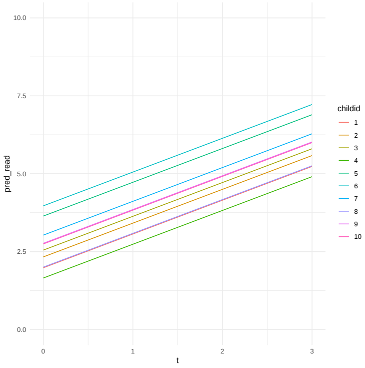
Models can be easily estimated using the `lmer()` function (`lme4` package)

```
> random_intercepts <- lmer(read ~ t + (1 | childid), data = reading)

> summary(random_intercepts)

Random effects:
 Groups   Name        Variance Std.Dev.
 childid  (Intercept) 0.7323   0.8557
 Residual             0.4225   0.6500
Number of obs: 884, groups:  childid, 221

Fixed effects:
            Estimate Std. Error t value
(Intercept)  2.71932    0.06820   39.87
t            1.08403    0.01955   55.44
```



$\rightsquigarrow$ $\bar{\alpha} = 2.72$

$\rightsquigarrow$ $var(\alpha_i) = 0.73$

(t is coded 0, 1, 2, 3)

Linear growth curve model

What if we want the rates of change to also vary by individuals?

$$y_{it} = \alpha_i + \beta_{1i} \cdot t_{it} + \epsilon_{it}$$

Note the subscript in both α_i *and* β_{1i}. **Both the intercept and the coefficients are allowed to vary individuals** i.

- ⇝ Known as a *Random slope linear growth model*
- ⇝ Captures individual variation in growth rate
- ⇝ α_i and β_{1i} are known as **growth parameters**

Basic linear growth curve model

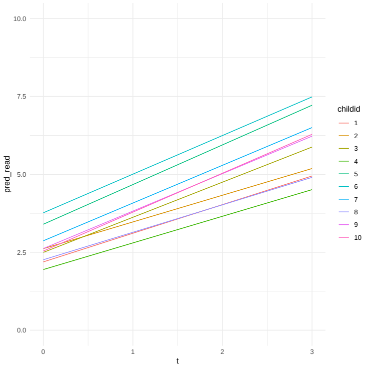
Models can be easily estimated using the `lmer()` function (`lme4` package)

```
> random_slope <- lmer(read ~ t + (t | childid), data = reading)

> summary(random_slope)

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 childid  (Intercept) 0.51927  0.7206
          t           0.06981  0.2642   0.15
 Residual             0.30645  0.5536
Number of obs: 884, groups:  childid, 221

Fixed effects:
            Estimate Std. Error t value
(Intercept)  2.71932    0.05762   47.19
t            1.08403    0.02436   44.51
```



- $\rightsquigarrow$ $\bar{\alpha} = 2.72$
- $\rightsquigarrow$ $var(\alpha_i) = 0.52$
- $\rightsquigarrow$ $var(\beta_{1i}) = 0.07$

(t is coded 0, 1, 2, 3)

Non-linear growth curve model

Quadratic curve growth models

$$y_{it} = \alpha_i + \beta_{1i} \cdot t_{it} + \beta_2 \cdot t^2 + \epsilon_{it}$$

Or even

$$y_{it} = \alpha_i + \beta_{1i} \cdot t_{it} + \beta_{2i} \cdot t^2 + \epsilon_{it}$$

- We can add a quadratic term t^2 to account for non-linear trajectories
- This quadratic term can or cannot vary by individual too (i.e., β_{2i})
    - Then α_i, β_{1i} and β_{2i} would be growth parameters
    - Costs in efficiency and convergence

Summary

- Longitudinal data can be modelled as multilevel
- Growth curve models estimate how individuals change over time
- Linear and quadratic terms allow for flexible trajectories
- Next: Hands-on examples and interpretation

Thank you!

thiago.oliveira@manchester.ac.uk

ThiagoROliveira.com

@oliveiratr.bsky.social