THE UNIVERSITY of EDINBURGH
Moray House School of Education and Sport

Analysing change: A brief introduction to Longitudinal Data Analysis

Patricio Troncoso

June 2025

Today’s programme

Time	Topic
10:00 – 10:10	Welcome and introduction
10:10 – 10:25	What is longitudinal data?
10:25 – 10:40	Analysing change
10:40 – 11:10	Practical 1
11:00 – 11:30	Multilevel model for change
11:30 – 11:45	Break
11:45 – 12:15	Latent Growth curve modelling (LGCM)
12:15 – 12:50	Practical 2
12:50 – 13:00	Wrap-up

What is longitudinal data?

What is longitudinal data?

- Quite simply:
  - Any data collected at the unit level about more than one occasion
- Also note that three conditions must be met:
  - Data is collected on **more than one units**
  - Units are **uniquely identified** over time
  - Data is collected at the **same level** of the units on more than one occasion
- Otherwise… data would be classified as cross-sectional
  - Also, if waves of longitudinal are analysed separately, then they are also cross-sectional

Longitudinal studies

- Longitudinal studies is a general term covering:
  - Cohort studies
  - Panel studies
  - Prospective studies
  - Follow-up studies
  - Growth studies
  - Repeated measures experiments
  - Event Histories
  - Pure and Mixed Longitudinal Designs
  - Accelerated Longitudinal Designs
- But not:
  - Time Series
  - Single Subject Designs

Longitudinal populations

- Longitudinal populations need to be defined **spatially** – just as cross-sectional populations do – but also **over time**.
- For example:, the Millennium Cohort Study population is a population of children defined as:

"all children born between 1 September 2000 and 31 August 2001 (for England and Wales), and between 24 November 2000 and 11 January 2002 (for Scotland and Northern Ireland), alive and living in the UK at age nine months, and eligible to receive Child Benefit at that age.

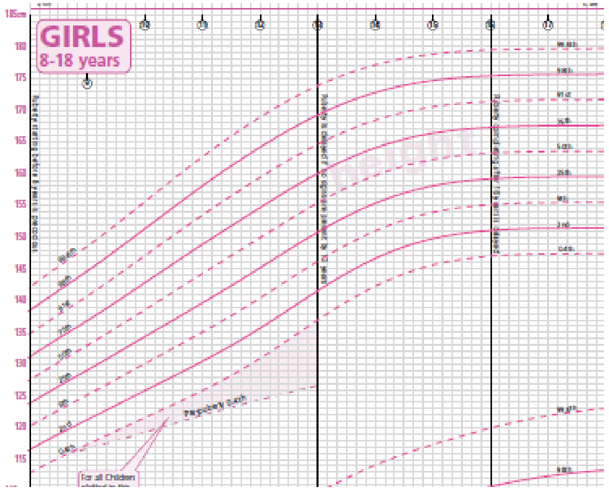
And, after nine months:

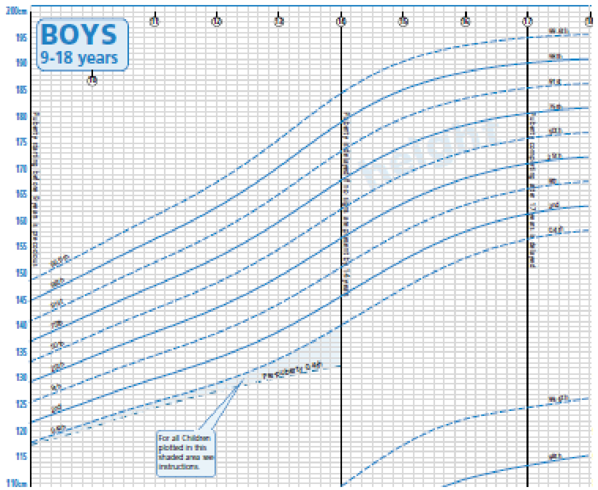
"for as long as they remain living in the UK at the time of sampling."

Key points to note: The definition is based on the population living in the UK at the first wave and the size of the population is allowed to decrease over time due to permanent emigration and deaths of cohort members.

Rationale for longitudinal studies

1. To study the **between or inter-individual (level 2)**, and **within or intra-individual (level 1)** relation of a characteristic with age or time.
   - Examples: What are the patterns of child growth or cognitive development in the early years of life?





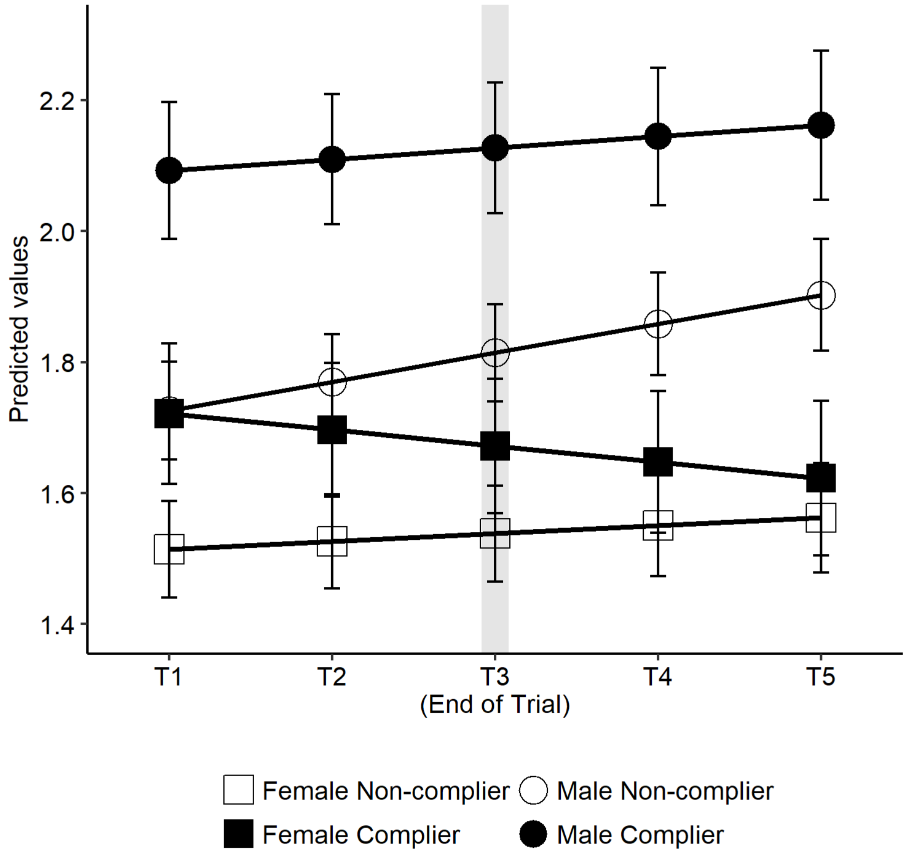
Source: WHO https://goo.gl/J8AVCe

Rationale for longitudinal studies (contd.)

2. To study the relation between **'earlier'** events and **'later'** outcomes
    - Example: Is getting a degree associated with better health in later life?

Rationale for longitudinal studies (contd.)

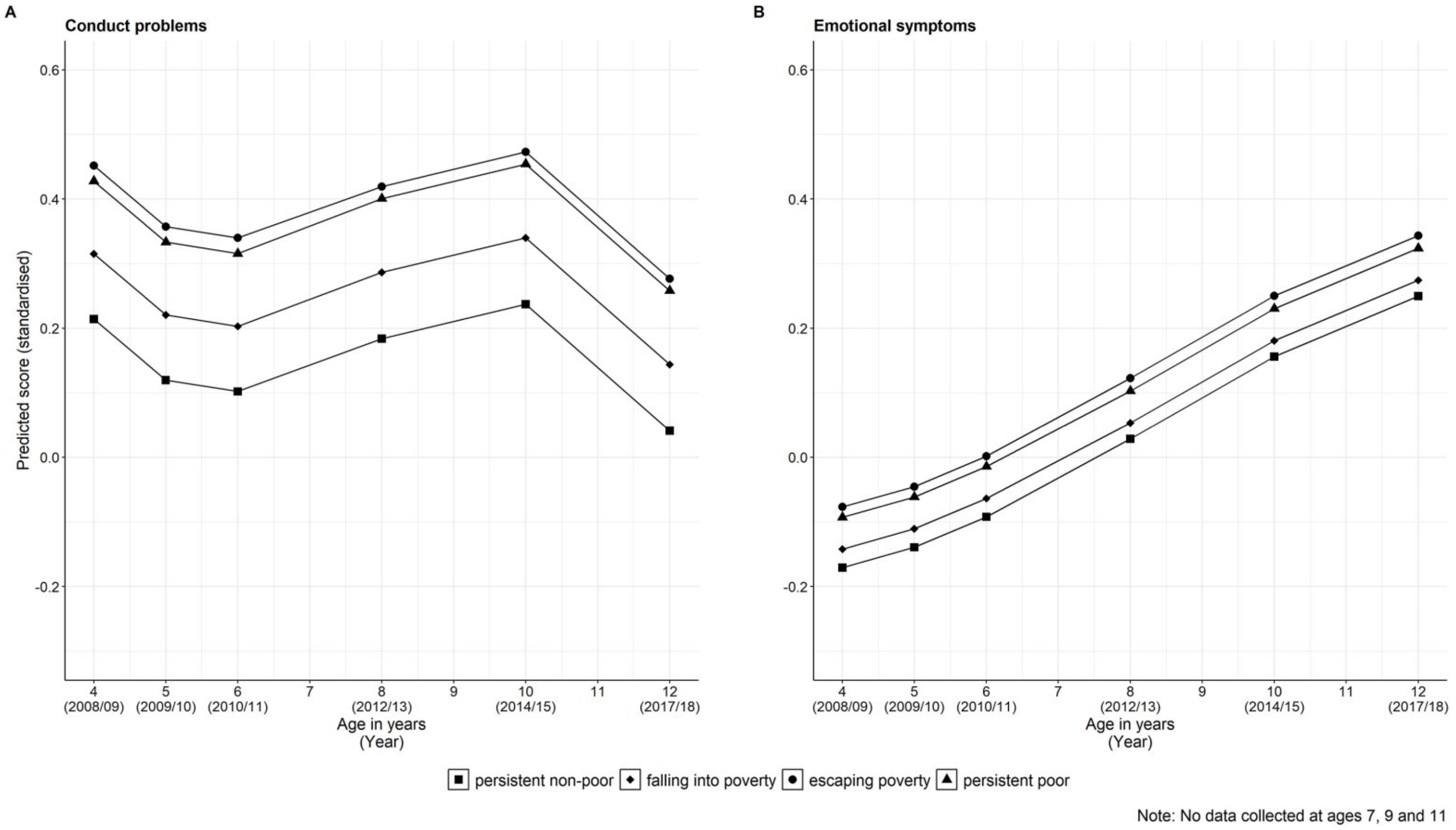
3. To evaluate the effects of social, educational or other types of interventions
   - Example: Are pupils who comply with the requirements of an intervention expected to have reduced disruptive behaviour compared to pupils under “usual practice”?



Source: Troncoso et al. (2024)

Rationale for longitudinal studies (contd.)

4. To compare other non-randomly formed groups over time (not formed as a result of an intervention)
   - Example: How do conduct problems and emotional symptoms in young people change over time according to lived experiences of poverty?



Source: Treanor and Troncoso (2022)

Rationale for longitudinal studies (contd.)

5. Causal models:

(i) To estimate a **postulated** causal model

- Example: the relation between unemployment and health.

(ii) To deduce a causal model from variables changing over time.

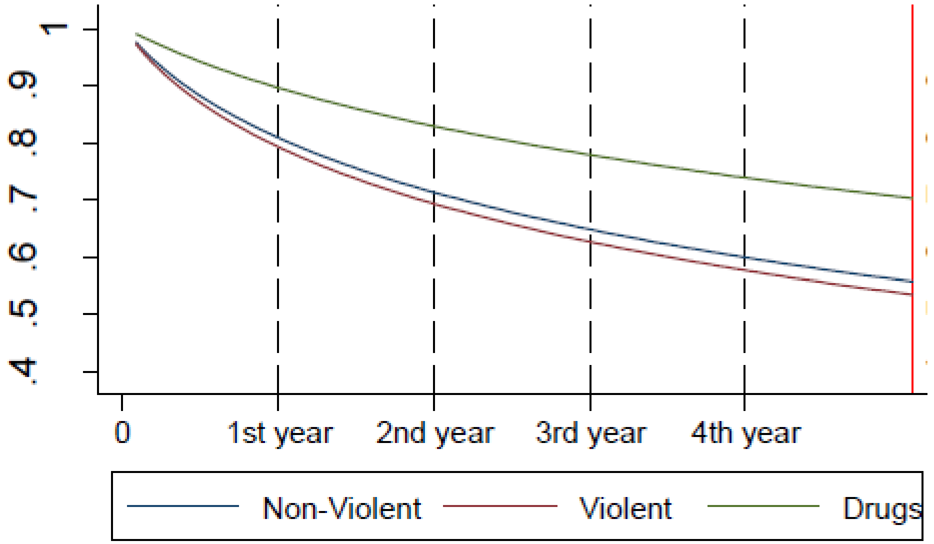
- Example: the relation between aggressive behaviour and exposure to media violence.

Rationale for longitudinal studies (contd.)

6. To measure, and model, the stability of a variable. Essentially, this would be the stability of individual differences over time
    - Example: **This one is for you to think!**
    - Why would it be important to know if a variable is stable over time?

Rationale for longitudinal studies (contd.)

7. To model the durations in states, and the transitions between states, as generated by event histories and by observation in continuous time.
   - Example: How long do ex-offenders take until reoffending according to type of offence?



Source: Morales-Gomez (2018)

Some advantages of longitudinal studies

1. Often the only way of measuring INDIVIDUAL CHANGE.
2. Can often separate AGE effects from COHORT effects on change.
3. Extends concepts by allowing DYNAMIC definitions of variables.
4. Collection of 'very' retrospective data can be avoided.

Some advantages of longitudinal studies (contd.)

5. Potentially valuable for causal analysis, allowing the effects of constant but unmeasured variables to be eliminated.
6. Avoids the difficulties associated with hidden state dependence.
7. Allows for control of survey errors, such as “telescoping”, using bounded recall and dependent interviewing.
8. Can provide good samples for cross-sectional studies.

There are always disadvantages…

1. Expensive, both in terms of data collection and administration.
2. Danger of sample loss over time, known as **sample attrition**, and likely to lead to bias.
3. Danger of **repeated measurement bias**, i.e. the reactive effect of prior measurement on current measures.
4. Results are not timely and tested hypotheses might be no longer interesting.
5. Methodologically complicated because of the need to change measuring instruments over time/age.
6. Statistically complicated and so data are often not analysed longitudinally.

Analysing change

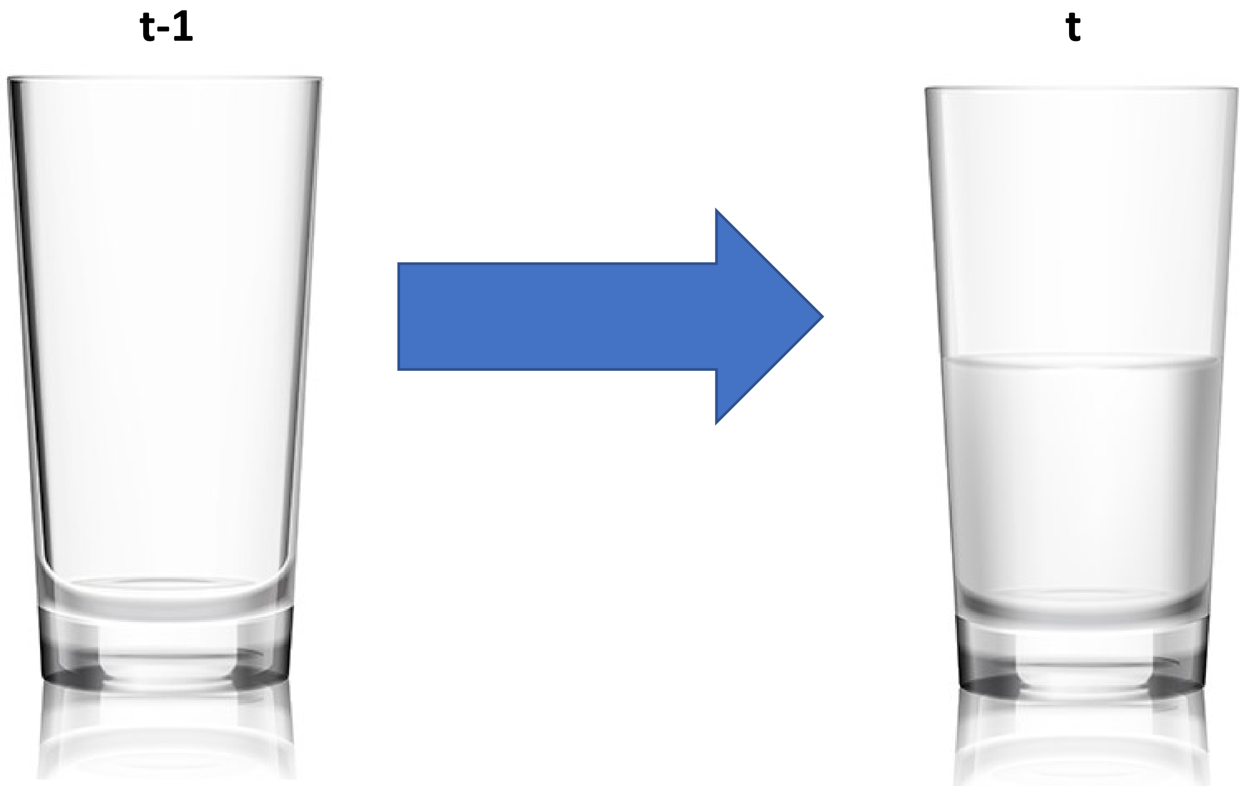
Is this glass half-empty or half-full?

The problem of cross-sectional data

- With only one snapshot at time t, we cannot draw meaningful conclusions without making further assumptions or extra information.
- What extra information would we need to establish causality?

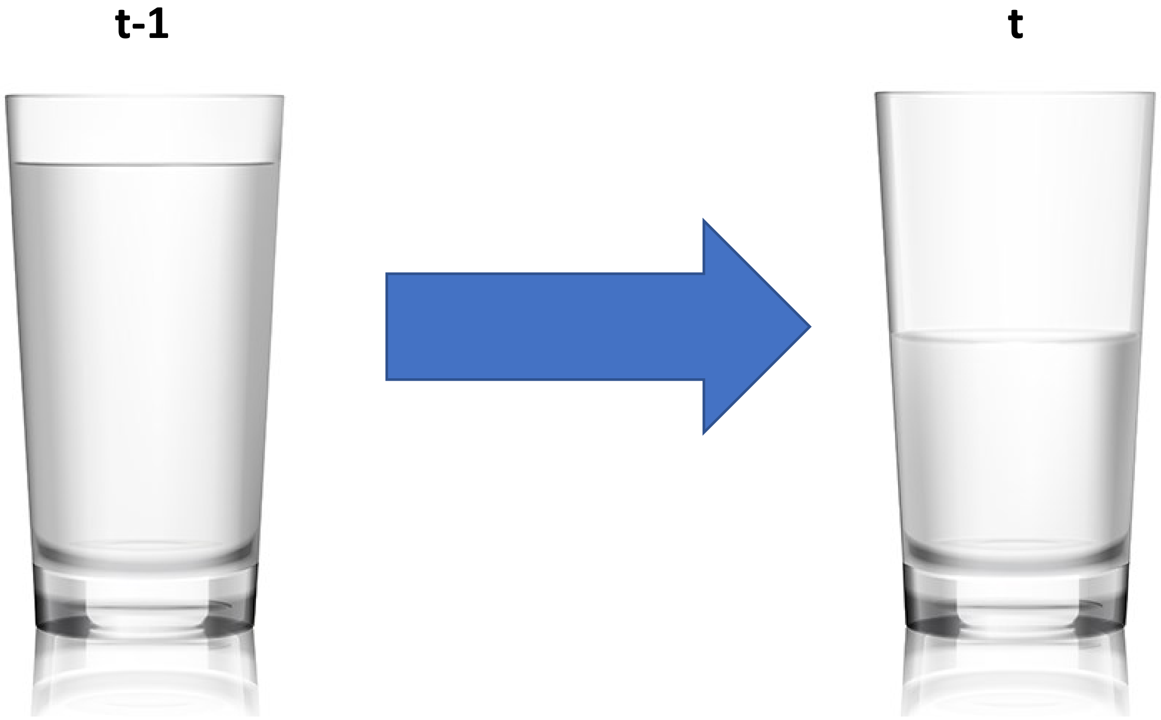
The problem of cross-sectional data

- The glass was being poured water in at time t-1, therefore the glass is half-full



The problem of cross-sectional data (contd)

- The glass was being drunk or poured out at time t-1, therefore the glass is half-empty



Dimensions of change

- *AGE* (A),
- *PERIOD/HISTORICALTIME* (P),
- *COHORT/GENERATION* (C) 'EFFECTS'
- Cohort (C) effects - change is only measurable at the **aggregate** level.
- Age (A) and period (P) effects - change can be measured for **individuals** and **aggregates** but cannot be separated at the individual level.

Dimensions of change (contd.)

- **Very important: C = P - A**
- so only **TWO** of the effects are identified in any one study.
- For further discussion of the identification problem, see Goldstein (1979)

Age, period and cohort effects

	AGE				
COHORT	0	10	20	30	40
1971					2011
1981				2011	
1991			2011		
2001		2011			
2011	2011				

How to read this table:
AGE (horizontal) by COHORT (vertical) by PERIOD (body of table, year in which we sample)

What sort of analysis can be conducted with this sample?

Age, period and cohort effects (contd.)

	AGE				
COHORT	0	10	20	30	40
1931					1971
1941				1971	1981
1951			1971	1981	1991
1961		1971	1981	1991	
1971	1971	1981	1991		
1981	1981	1991			
1991	1991				

How to read this table:
AGE (horizontal) by COHORT (vertical) by PERIOD (body of table, year in which we sample)

What sort of analysis can be conducted with the sample circled in red?

What sort of analysis can be conducted with sample circled in blue?

Age, period and cohort effects (contd.)

	AGE				
COHORT	0	10	20	30	40
1931					1971
1941				1971	1981
1951			1971	1981	1991
1961		1971	1981	1991	
1971	1971	1981	1991		
1981	1981	1991			
1991	1991				

How to read this table:
AGE (horizontal) by COHORT (vertical) by PERIOD (body of table, year in which we sample)

What sort of analysis can be conducted with the samples circled in red?

Accelerated longitudinal design

An ALD is a structured multiple cohort design that takes multiple single cohorts, each one starting at different age.

	AGE			
COHORT	0	10	20	30
1981			2001	2011
1991		2001	2011	
2001	2001	2011		

How long is the data collection period?

How many cohorts are there?

What is the age range?

What does longitudinal data look like?

- Quick answer: It depends…
- Slightly longer answer:
  - If two occasions, probably wide format
  - If more than two occasions, probably long format
- Let’s look at some examples

Repeated Measures Data (wide format)

Subject	Occasion 1	2	3	4
1	y_{11}	y_{21}		y_{41}
2	y_{12}	y_{22}	y_{32}	y_{42}
3	y_{13}		y_{33}	y_{43}

y_{ij} is the response at occasion *i* for individual *j*

View as a two-level structure (responses within individuals)

This example table is in the “wide” format

Repeated Measures Data (long format)

Occasion	**Subject**	**y**	**x1**	**x2**
1	*1*	*10*	*2*	*5*
2	*1*	*12*	*3*	*5*
3	*1*	*14*	*2*	*5*
1	*2*	*8*	*0*	*3*
2	*2*	*10*	*0*	*3*
3	*2*	*12*	*0*	*3*

This is data structured in the "long" format, i.e. each measurement occasion is a row in the data grid and individuals are another variable.

Longitudinal Descriptive Statistics

Practical One

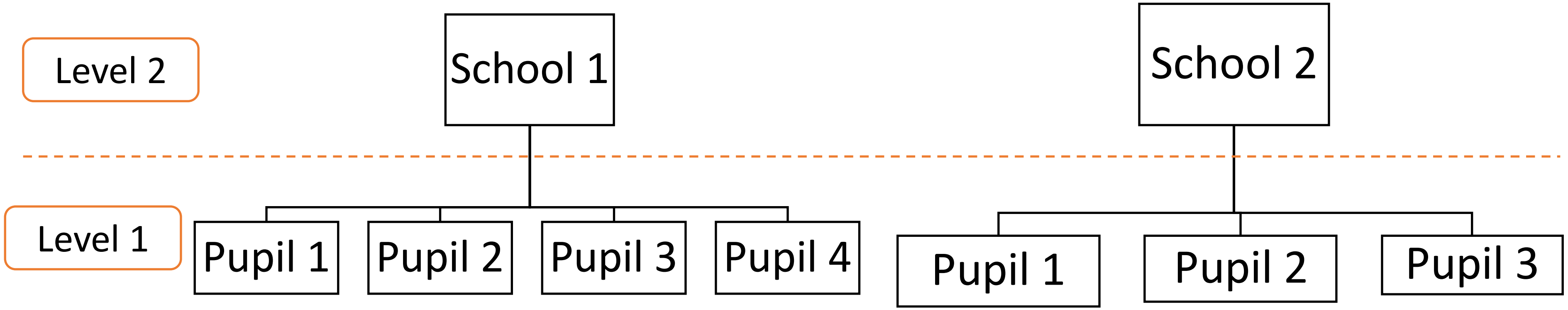
The Multilevel Model for change

How do we measure (individual) change?

- Repeated measures on a set of individuals
  - (Which are representative of a population)

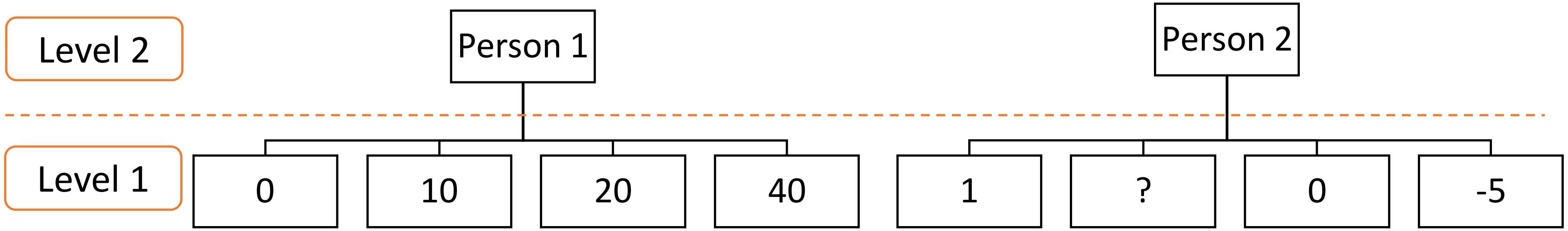
Subject	Score 1	Score 2	Score 3	Score 4
Person 1	0	10	20	40
Person 2	1		0	-5

Data structures



- This would be a typical hierarchical structure of individuals nested within clusters
  - A standard 2-level model

Data structures (contd.)



- This is still a standard 2-level structure:
  - What is the difference with the previous diagram?

Multilevel models

- A standard empty 2-level model has the following form:
  - This could be pupils (i) nested within schools (j)

$$\underbrace{y_{ij}}_{\text{Outcome}} = \underbrace{\beta_0}_{\text{Overall mean}} + \underbrace{u_{0j}}_{\text{Cluster-specific residual}} + \underbrace{e_{ij}}_{\text{Individual heterogeneity}}$$

This is also known as an unconditional means model

Multilevel models (contd.)

- With the estimates of the empty model, we can evaluate the relative magnitude of the variance components:

$$\rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2}$$

This is known as Variance Partition Coefficient (VPC) or Intraclass Correlation (ICC)

Multilevel models (contd.)

- Then, we could add a covariate x (at level 1), which would render this a random intercepts model
  - This model would have the following form:

$$\underbrace{\boxed{y_{ij}}}_{\text{Outcome}} = \underbrace{\boxed{\beta_0}}_{\text{Overall mean}} + \underbrace{\boxed{\beta_1 x_{ij}}}_{\text{Expected effect of covariate}} + \underbrace{\boxed{u_{0j}}}_{\text{Cluster-specific residual}} + \underbrace{\boxed{e_{ij}}}_{\text{Individual heterogeneity}}$$

Multilevel model for repeated measures

- A longitudinal model for occasions (i) nested within individuals (j) has the following form:
  - Just like MLM, but now the terms have different conceptual meanings:

$$\underbrace{y_{ij}}_{\text{Outcome}} = \underbrace{\beta_0}_{\text{Overall mean}} + \underbrace{\beta_1 t_{ij}}_{\text{Growth rate}} + \underbrace{u_{0j}}_{\text{Residual between individuals}} + \underbrace{e_{ij}}_{\text{Residual within individuals}}$$

Note: Time needs to be centred around a particular occasion

A random intercepts MLM for repeated measures?

- In the previous 2 slides, both equations are random intercepts models:
  - which assume that β_1 does not vary across level 2 units
- This is unrealistic for repeated measures
  - Why?

Assuming varying growth rates

- We can allow the growth rate β_1 to vary across individuals (level 2):

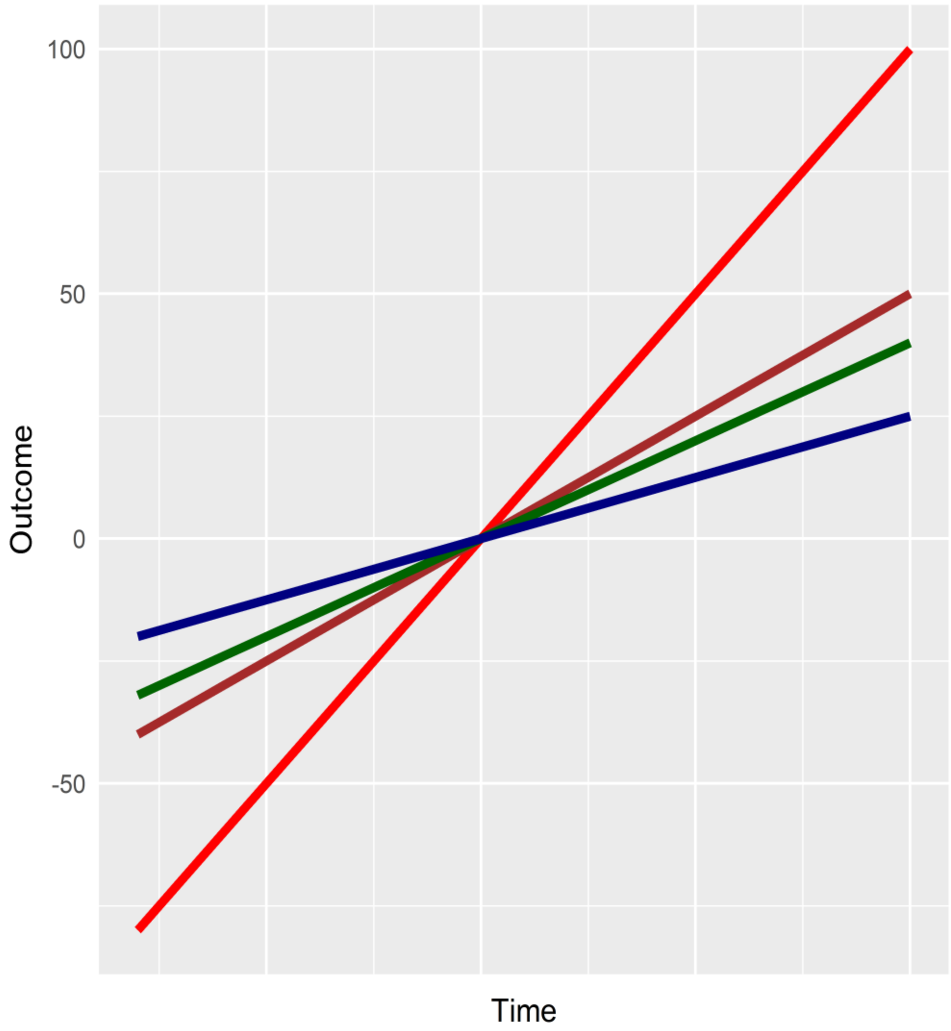
$$y_{ij} = \beta_{0j} + \beta_{1j}t_{ij} + u_{0j} + e_{ij}$$
$$\beta_{0j} = \beta_0 + u_{0j}$$
$$\boxed{\beta_{1j} = \beta_1 + u_{1j}}$$

This allows the growth rate to vary across individuals

In MLM jargon this is a random slopes model

In LDA jargon, this is an unconditional growth model

Linear growth curve

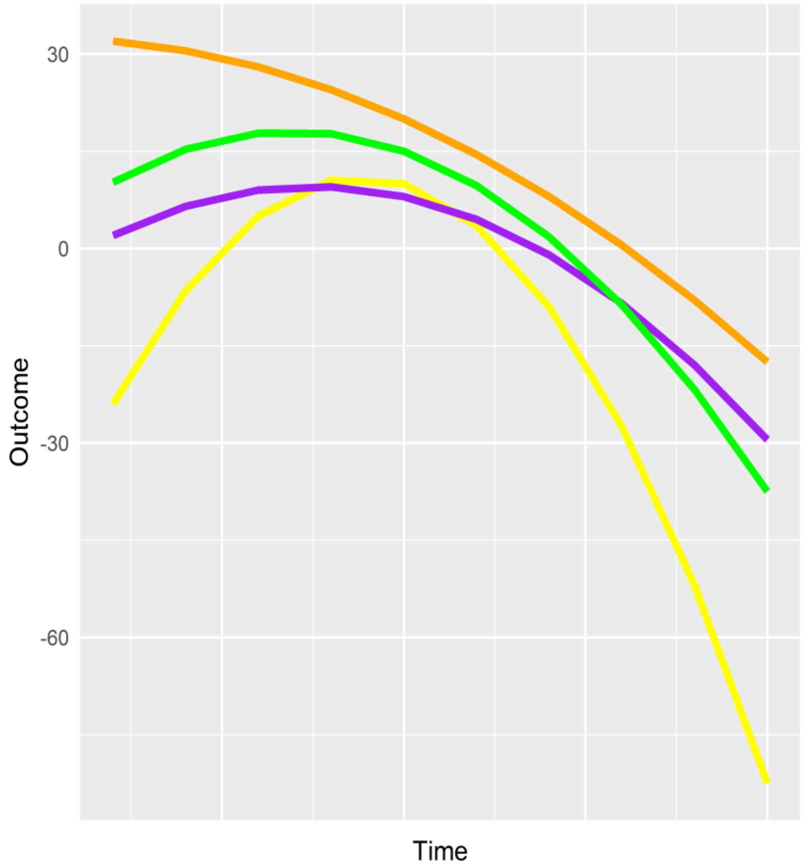


- The model described in the previous slide would look this graph
- Expected individual trajectories over time are linear and do not have the same slope

Nonlinear growth

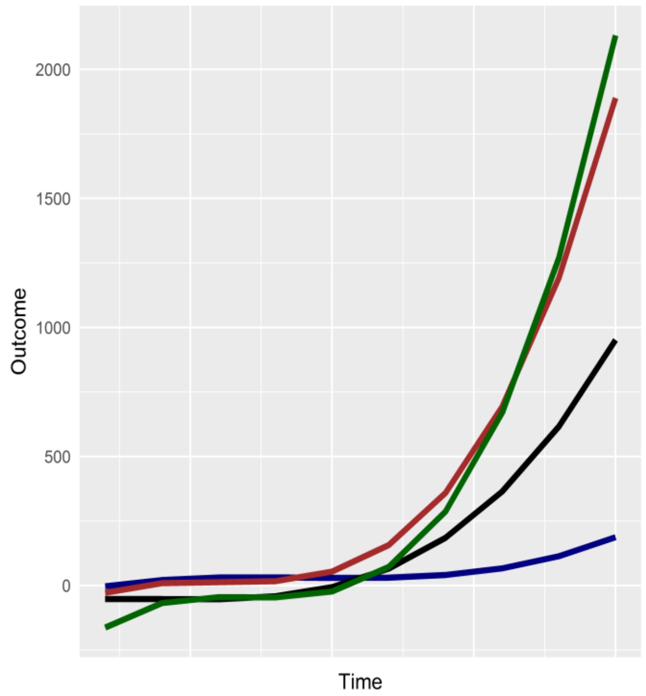
- Sometimes linear trajectories are not realistic and do not fit the data well.
  - Adding polynomials can help

$$y_{ij} = \beta_{0j} + \beta_{1j}t_{ij} + \beta_2 t_{ij}^2 + u_{0j} + e_{ij}$$



Nonlinear growth (contd.)

$$y_{ij} = \beta_{0j} + \beta_{1j}t_{ij} + \beta_2 t_{ij}^2 + \beta_3 t_{ij}^3 + u_{0j} + e_{ij}$$



- This can help to control for floor and ceiling effects
  - In the case of variables with maxima and minima

Nonlinear growth (contd.)

- We can specify even more flexible models, allowing for the nonlinear growth rates to vary across individuals:

$$y_{ij} = \beta_{0j} + \beta_{1j} t_{ij} + \beta_{2j} t_{ij}^2 + \beta_{3j} t_{ij}^3 + u_{0j} + e_{ij}$$

- But first of all: Why would we need/want to do this?

A step-by-step growth curve analysis

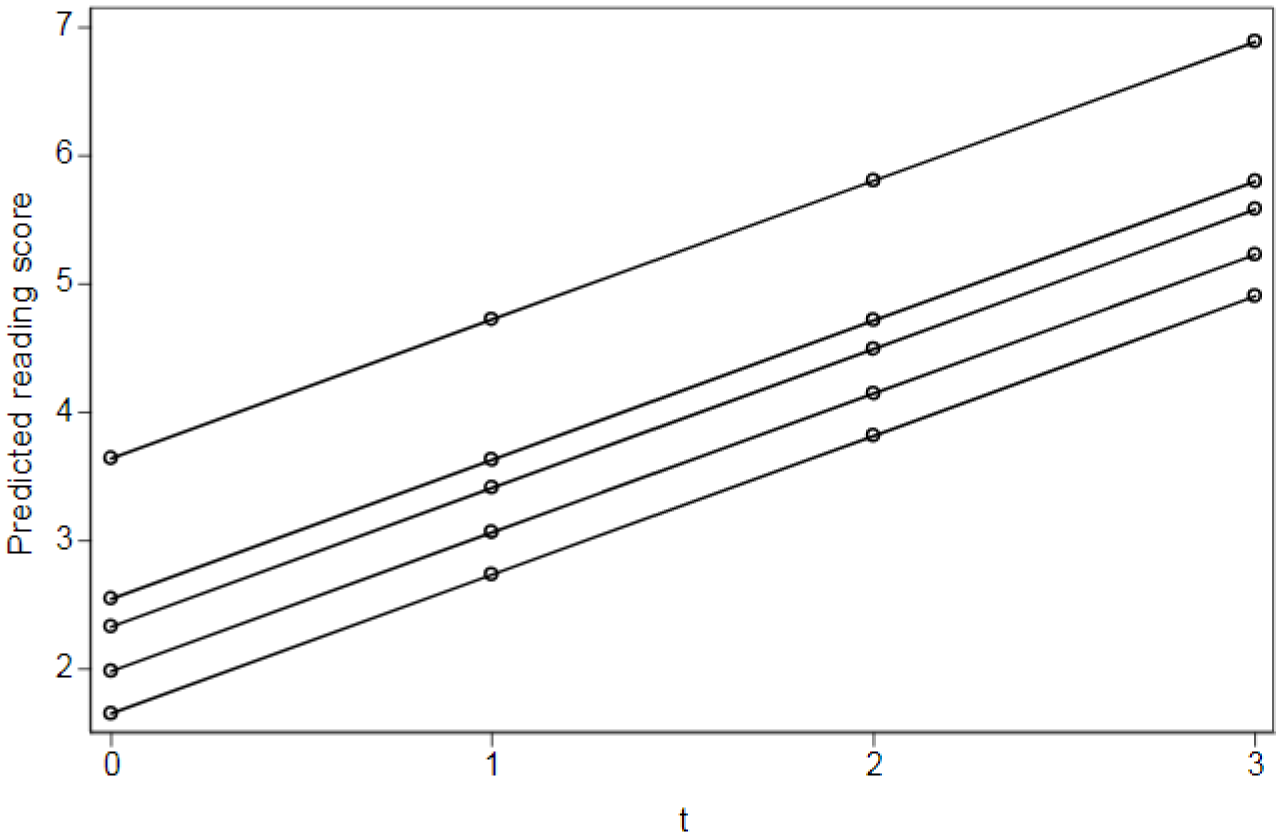
Random intercept growth model for reading progress

Parameter	Estimate	St. Error
Constant (β_0)	2.72	0.07
t (β_1)	1.08	0.02
Between-individual variance (σ^2_{u0})	0.73	0.08
Within-individual variance (σ^2_e)	0.42	0.02
- log-likelihood	1101.3	

Source: Steele, F. (2014). Multilevel Modelling of Repeated Measures Data. LEMMA VLE Module 15, 1-62. (http://www.bristol.ac.uk/cmm/learning/course.html).

A step-by-step growth curve analysis (contd.)

Fitted reading trajectories for 5 selected children



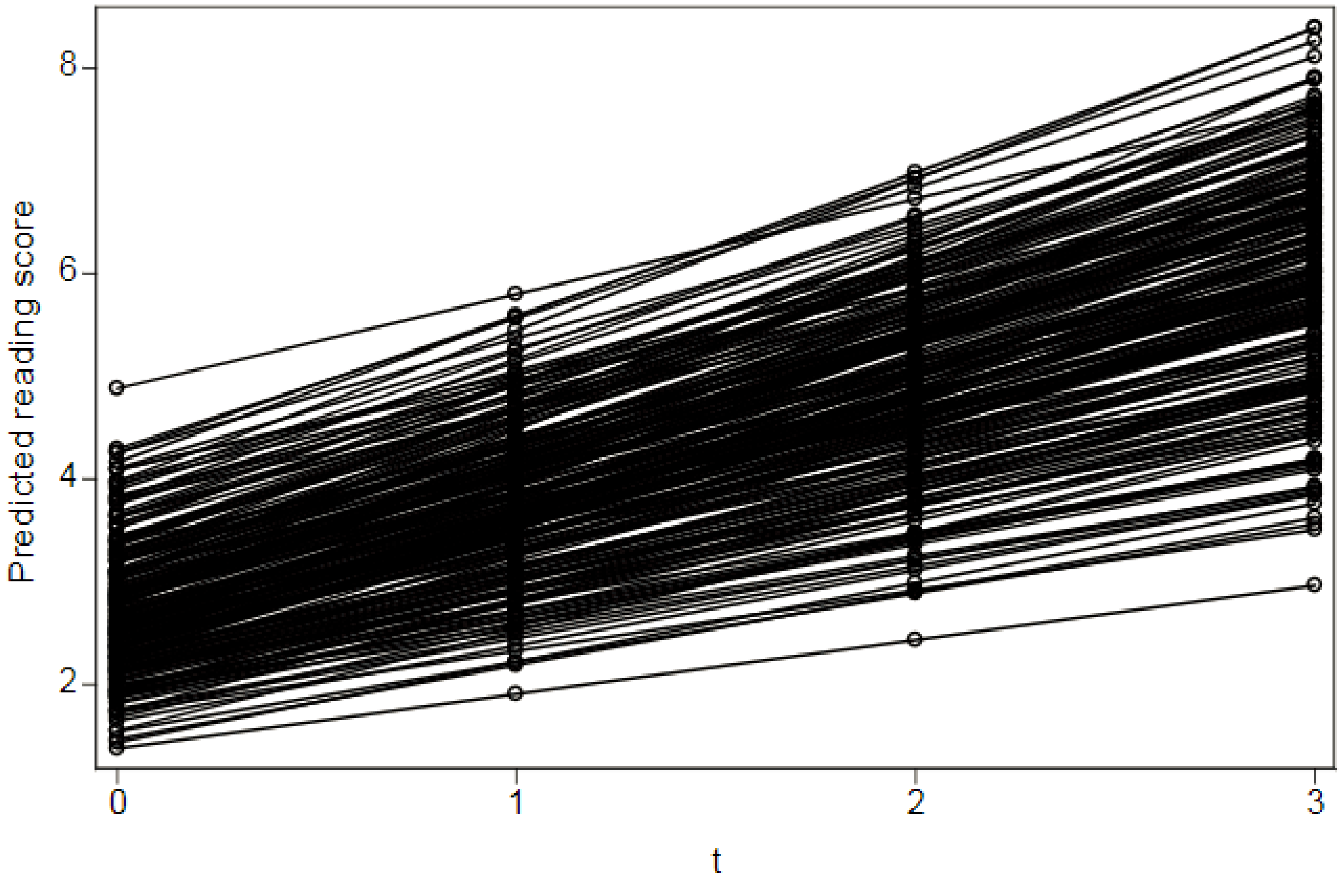
A step-by-step growth curve analysis (contd.)

Random slope growth model for reading progress

Parameter	**Estimate**	**St. Error**
Constant (β_0)	2.72	0.06
t (β_1)	1.08	0.02
Between-individual intercept variance (σ^2_{u0})	0.52	0.07
Between-individual slope variance (σ^2_{u1})	0.07	0.01
Between-individual intercept-slope covariance (σ_{u01})	0.03	0.02
Within-individual variance (σ^2_e)	0.31	0.02
-log-likelihood	1059.5	

A step-by-step growth curve analysis (contd.)

Fitted reading trajectories for all children in the sample



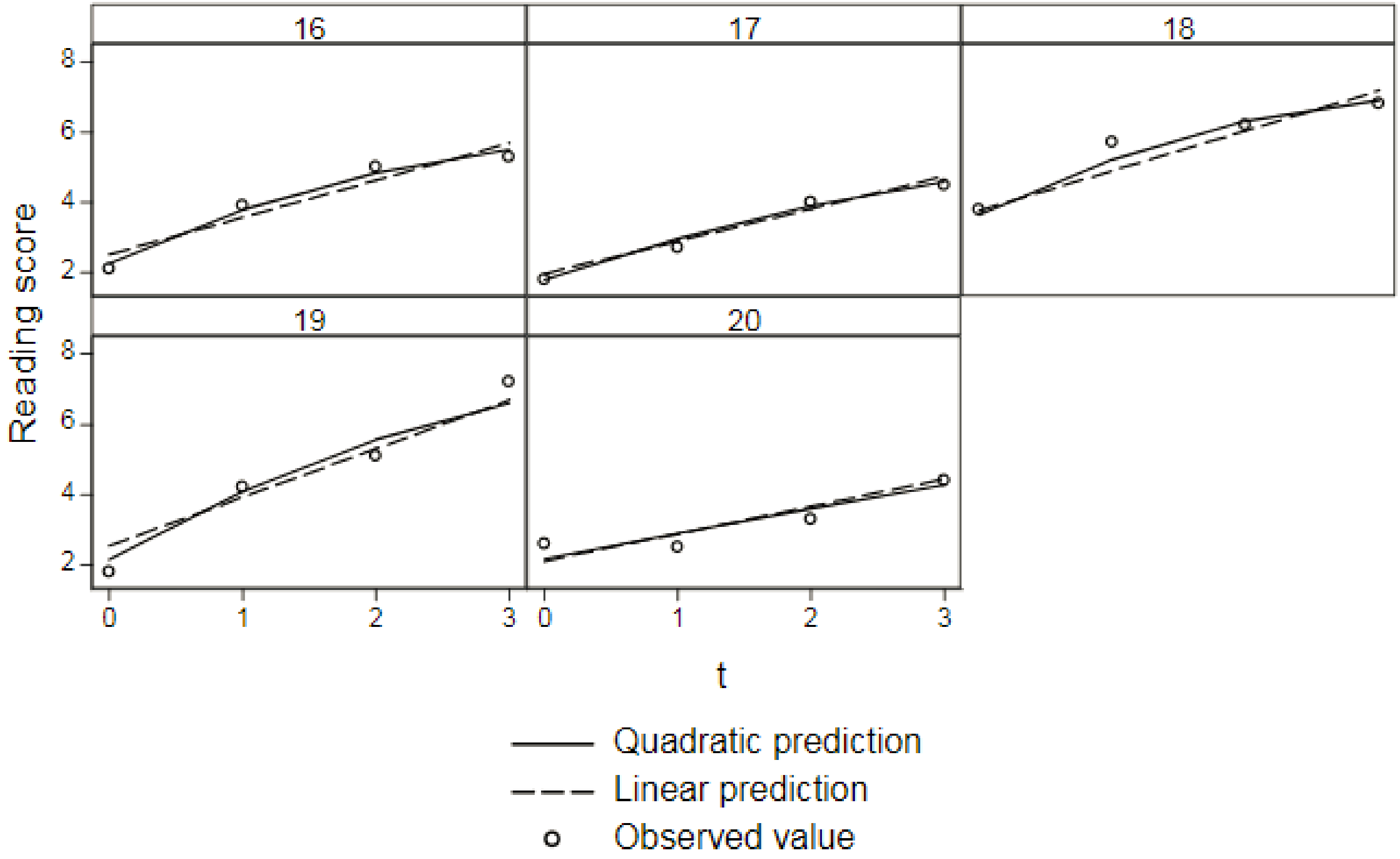
A step-by-step growth curve analysis (contd.)

Quadratic growth model for reading progress

Parameter	**Estimate**	**St. Error**
Constant (β_0)	2.53	0.06
t (β_1)	1.64	0.06
t^2 (β_2)	-0.19	0.02
Between-individual variances/covariances		
Intercept variance (σ^2_{u0})	0.57	0.07
t variance (σ^2_{u1})	0.36	0.09
t^2 variance (σ^2_{u2})	0.02	0.01
Intercept - t covariance (σ_{u01})	-0.02	0.06
Intercept - t^2 covariance (σ_{u02})	-0.002	0.02
t - t^2 covariance (σ_{u12})	-0.07	0.03
Within-individual variance (σ^2_e)	0.20	0.02
-log-likelihood	994.0	

A step-by-step growth curve analysis (contd.)

Fitted reading trajectories for 5 children from random slope linear and quadratic growth models



Adding further explanatory variables

- Individual variables are now level 2 variables
  - They can be added just like any other variable
- Time-varying predictors can also be added
  - This is simply a level 1 variable, just like time in the previous unconditional growth models
- Variables can interact within and between levels
- Not exactly explanatory variables, but:
  - The MLM specification allows for extensions to further levels above the individual
  - When would this be of interest?

Model comparison

- Depending on the estimator you use:
  - If maximum likelihood:
    - Likelihood ratio
      $$LR = -2\ logL_1 - (-2logL_2)$$
    - Then compare to a χ^2 distribution with df equals the number of extra parameters
    - Alternatively compare AIC (Akaike Information Criterion) between models: the smaller the better
  - If Bayesian estimation:
    - Compare DIC (Deviance Information Criterion) between models: the smaller the better

Assumptions

- Again, just like in MLM:
  - Normality of residuals at each level
  - Homoscedasticity: equal variances of residuals at each level across values of every predictor

Latent Growth Curve Models

Growth curve modelling

- Individuals are assumed to have a ‘trajectory’ over time with regard to their observed responses
- The simplest model assumes linear trajectories, but we can hypothesise and test trajectories of almost any form as long as we have enough data (time-points)

Analysing longitudinal data

- Two approaches to fitting/estimating Growth Curve Models
  - Multilevel or Mixed modelling
    - ‘Univariate’ approach: The outcome is treated as one variable measured at several different times. Within-person correlations are handled by treating the data as nested (e.g. occasions nested within individuals gives two ‘levels’) and including random effects.
    - Known as “Growth Curve Modelling”.
    - You can do this in R (lme4, nlme and other packages), Stata, SPSS, MLwiN

Analysing longitudinal data (contd.)

- Two approaches to fitting/estimating Growth Curve Models
  - Structural Equation Modelling
    - ‘Multivariate’ approach: The outcomes over time are viewed as several different variables, one for each time point. Within-person correlations are handled by assuming the presence of latent variables (i.e. unobserved causes), called growth factors.
    - Known as **“Latent** Growth Curve Modelling”.
    - This can be done with the R package “lavaan”, Mplus, Stata, and others

Univariate vs. Multivariate

Univariate i.e. Multilevel

Growth curve modelling

person	time	Y
1	1	9
1	2	7
1	3	10
2	1	6
2	2	7
2	3	9

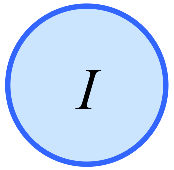
Multivariate i.e. SEM

Latent Growth Curve Modelling

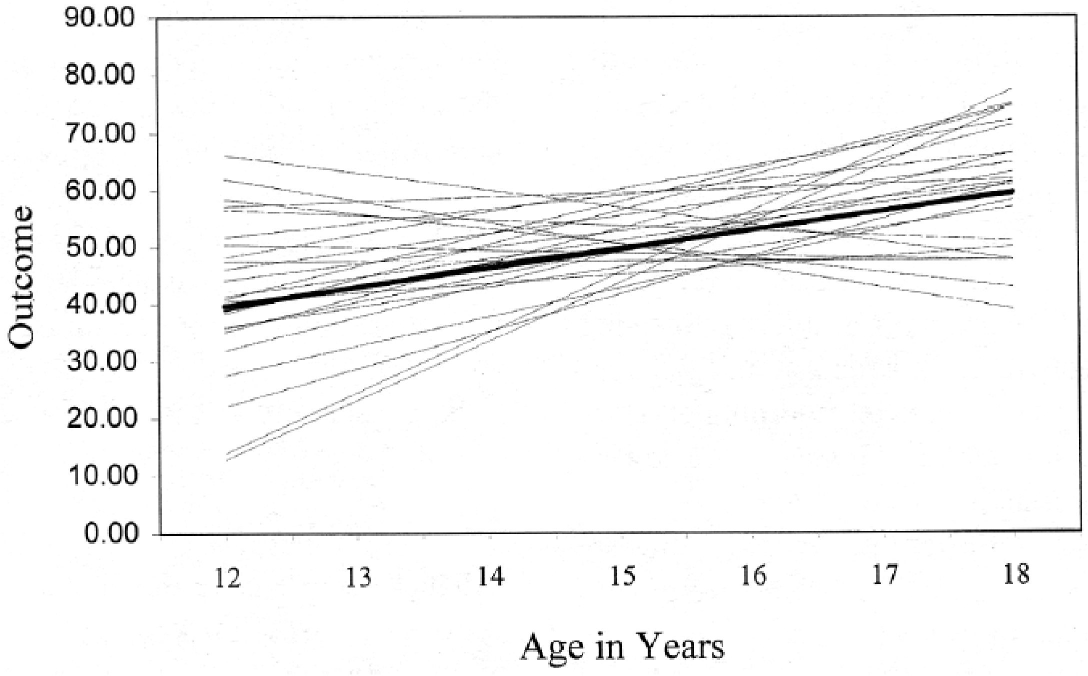
person	Y1	Y2	Y3
1	9	7	10
2	6	7	9

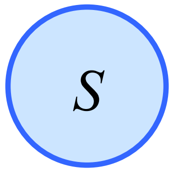
Long format vs. wide format

Latent Growth Factors



Intercept factor – two parameters:
Mean: Average of the outcome variable at time 1.
Variance: Individuals' variation around this average.





Slope factor – two parameters:
Mean: Average difference from one time to the next.
Variance: Individuals' variation around this average.

Source: Mehta & West (2000). Psych. Meth., 5(1):23-43

Parameterisation of the LGCM

$$Y_{tj} = I_j + \lambda_t S_j + \varepsilon_{tj}$$

Y_{tj} = response of person *j* at time *t*

I_j = intercept latent growth factor

S_j = slope latent growth factor

λ_t = loading for time *t*

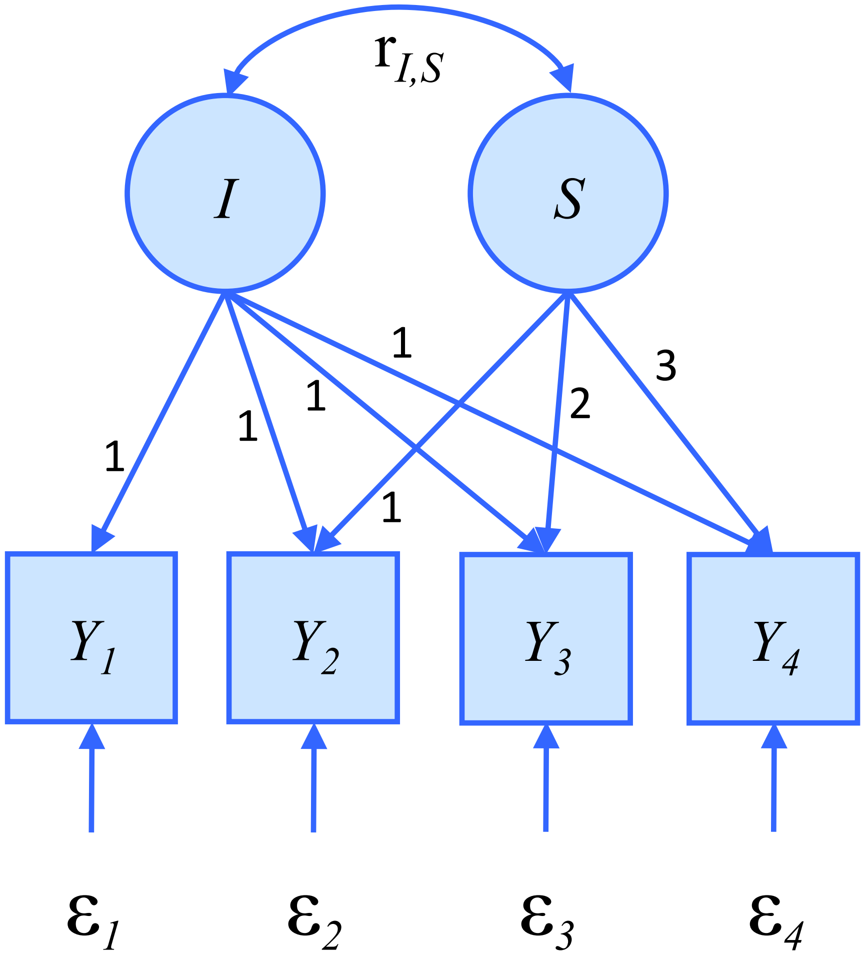
ε_{tj} = Residual for person *j* at time *t*

Parameterisation of the LGCM

$$Y_{tj} = I_j + \lambda_t S_j + \varepsilon_{tj}$$

- **We** specify the loadings for the Intercept and Slope factors
  - These describe/specify the form of the growth curve
  - e.g. linear change over time
- We then **estimate** the mean and variance of I and S (and their correlation, and the residuals)

Parameterisation of the LGCM



$$Y_1 = I + \epsilon_1$$

$$Y_2 = I + S + \epsilon_2$$

$$Y_3 = I + 2S + \epsilon_3$$

$$Y_4 = I + 3S + \epsilon_4$$

Item scores (Y) are a function of the latent growth factors, plus occasion-specific error.

Interpretation of parameters

$$Y_{tj} = I_j + \lambda_t S_j + \varepsilon_{tj}$$

E(I) = Average Y score at time 1

Var(I) = Variance in E(I)

E(S) = Average change in Y score between time points

Var(S) = Variance in E(S)

$r_{I,S}$ = Correlation between I and S

Fitting a LGCM in R

- lavaan syntax:

```
lgcm <- ' i =~ 1*y1 + 1*y2 + 1*y3 + 1*y4
          s =~ 0*y1 + 1*y2 + 2*y3 + 3*y4 '
```

  - The code above is a linear growth curve

```
fit2 <- growth(lgcm, data=data)
```

  - The code above calls the function “growth” to run the model
  - **This would be an unconditional linear growth model**
  - Here we have equally spaced occasions. Slope indicators (y1-y4) are multiplied by 0-3
    - If measures were taken at varying intervals, e.g.:
      - `s =~ 0*y1 + 1*y2 + 3*y3 + 7*y4`

Model fit

- The LGCM is a *model-based description* of the data. How good a description?
- We can use the standard goodness of fit indices to evaluate the model's global properties

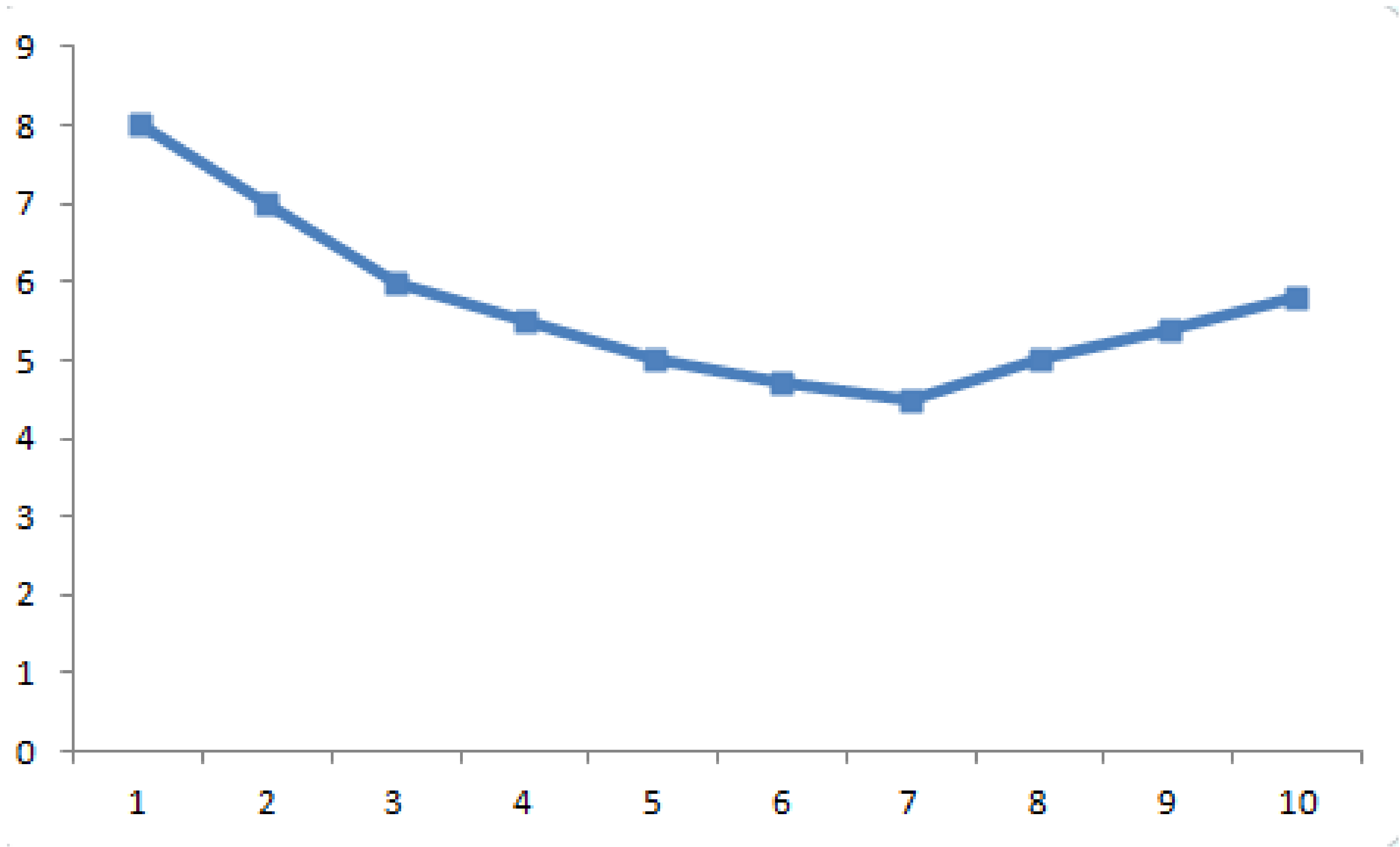
Model fit summary

- Hu & Bentler (1999) suggest using two indices, with the following cut offs:
    1. SRMR (< 0.08 = "good" fit), plus
    2. Either RMSEA (< 0.06) or CFI (> 0.95).
- AIC, BIC are also available
    - Reminder: the smaller the better
- But we shouldn't just blindly use cut-off criteria as oracles of 'truth'.
    - Theory should be the first and foremost decider.

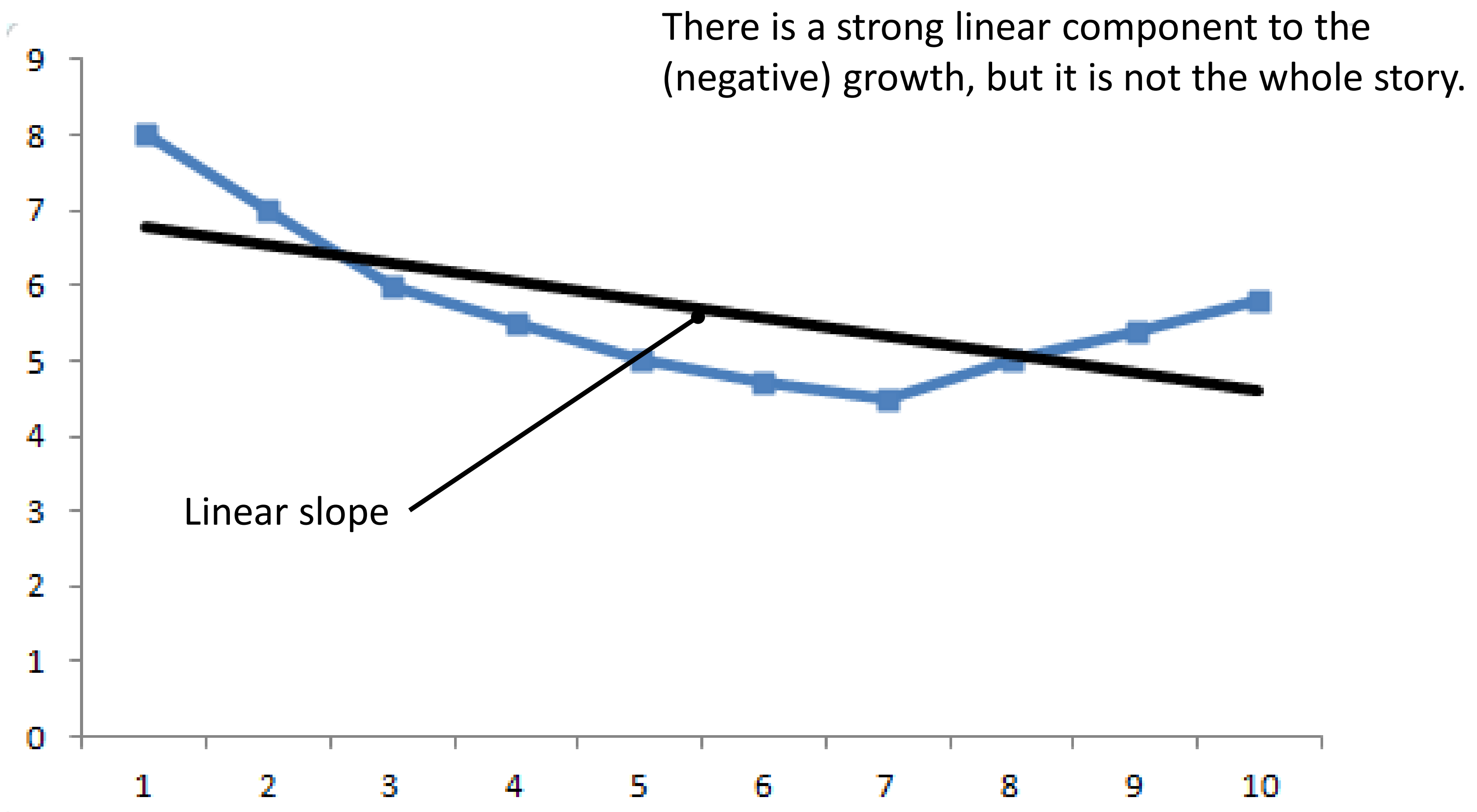
What if my LGCM doesn’t fit well?

- Poor fit implies something isn’t quite right with the model. Often this is because the underlying assumptions of the model are not being met.
- One big assumption in the current model:
  - Linear growth over time!

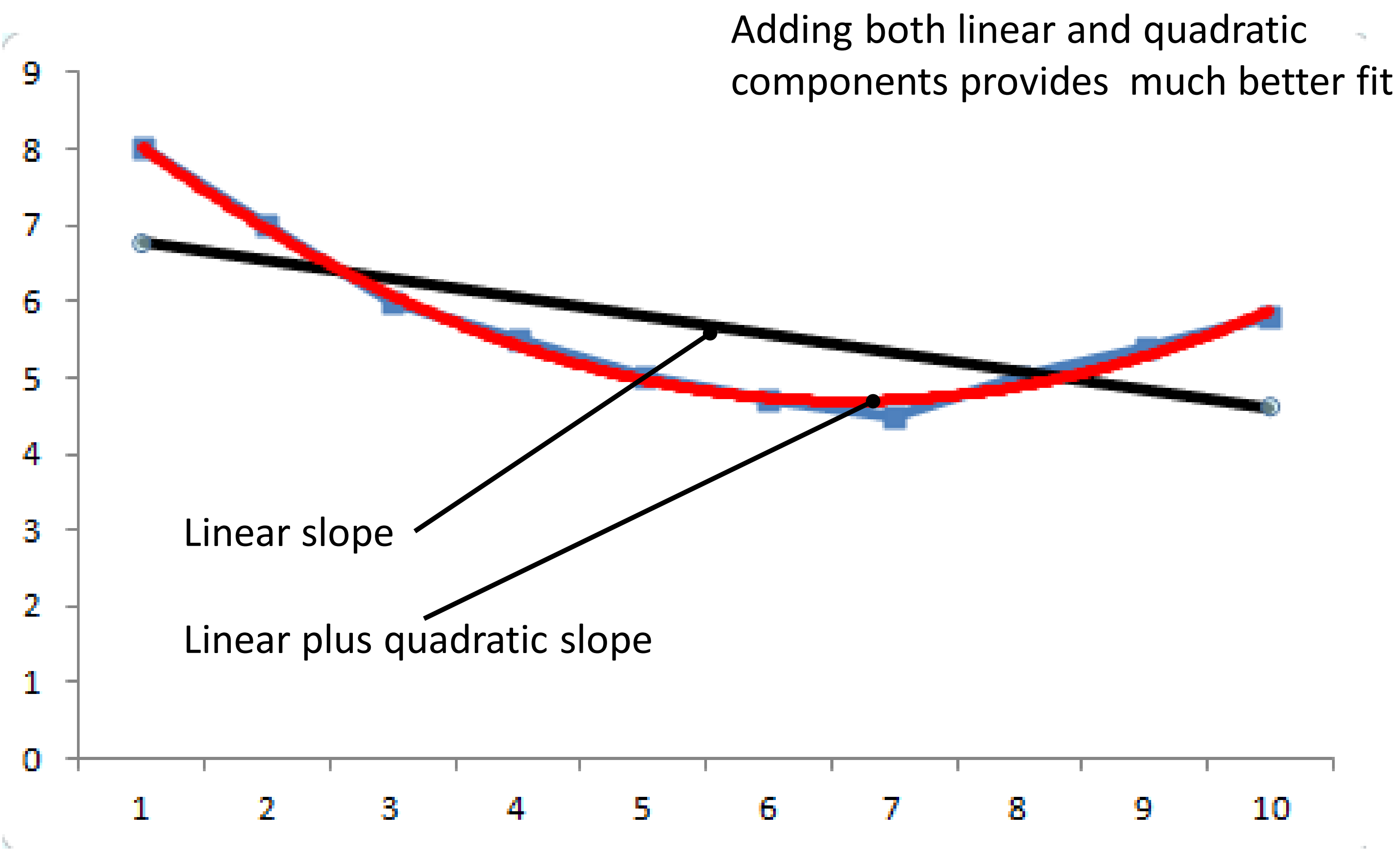
Non-linear growth



Non-linear growth



Non-linear growth



Fitting a quadratic LGCM in R

- lavaan syntax:

```
quad_lgcm <- ' i =~ 1*y1 + 1*y2 + 1*y3 + 1*y4
               s =~ 0*y1 + 1*y2 + 2*y3 + 3*y4
               q =~ 0*y1 + 1*y2 + 4*y3 + 9*y4‘
```

  - The code above is a quadratic growth curve

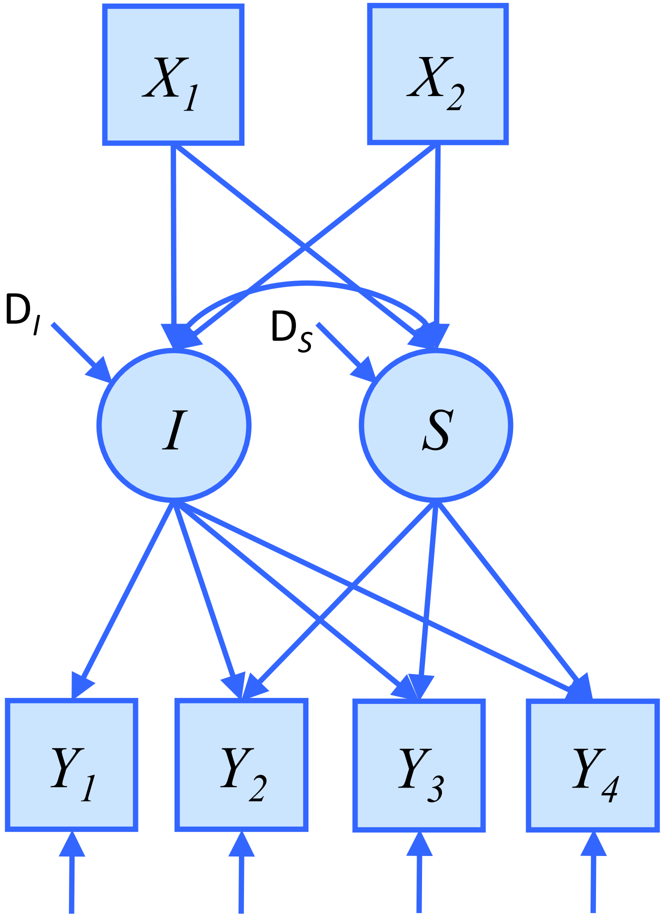
```
fit2 <- growth(quad_lgcm, data=data)
```

  - The code above calls the function “growth” to run the model

Adding predictors

- So far, the LGCM is just a model-based *description* of change in the outcome variable over time.
- Usually we want to see what variables are associated with that change – we need to add in some predictors.
- Some predictors are stable over time, or time-invariant
  - Date of birth (cohort), sex
- Some predictors are changing over time, or time-varying

Predictors of the growth factors



I and S are now outcome variables (i.e. ‘endogenous’), regressed upon observed predictors (Xs).

Their residual variances are termed ‘disturbances’.

This is a parsimonious way to model the effects of covariates that don’t change over time.

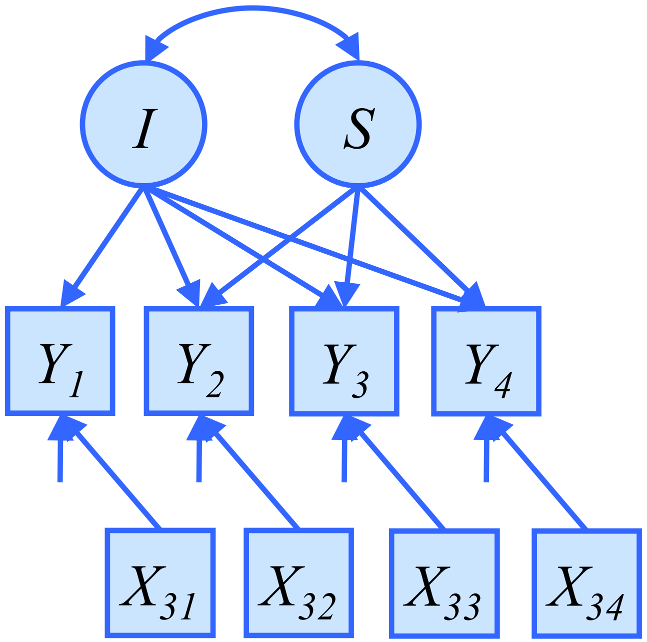
Fitting a LGCM in R with time invariant predictors

- lavaan syntax:

```
lgcm2 <- ' i =~ 1*y1 + 1*y2 + 1*y3 + 1*y4
           s =~ 0*y1 + 1*y2 + 2*y3 + 3*y4
           i ~ x1 + x2
           s ~ x1 + x2 '
```

- The effects of x1 and x2 on the intercept and the slope are estimated

Time-varying predictors



X_{31} – X_{34} represent four observations of a predictor that varies over time. Each Y is regressed on its associated X.

Note that, if using the multilevel GCM approach, these four regressions are usually assumed equal if no interactions with time are specified, but vary by default here.

Fitting a LGCM in R with time varying predictors

- lavaan syntax:

```
lgcm3 <- ' i =~ 1*y1 + 1*y2 + 1*y3 + 1*y4
           s =~ 0*y1 + 1*y2 + 2*y3 + 3*y4
           i ~ x1 + x2
           s ~ x1 + x2
           y1 ~ c1
           y2 ~ c2
           y3 ~ c3
           y4 ~ c4'
```

- The effects of x1 and x2 on the intercept and the slope are estimated

Some take-home points about LGCM and GCM

- The SEM framework can be useful for analysing longitudinal data
  - It allows great flexibility
- The SEM framework is a different specification of the same problem that the MLM for change analyses
  - Models fitted from both approaches are equivalent

Fitting a LGCM

Practical 2

References

- Singer, J., & Willett, J. (2003). Applied longitudinal data analysis: modeling change and event occurrence. Oxford University Press.
- Long, J. D. (2011). Longitudinal Data Analysis for the Behavioral Sciences Using R. Thousand Oaks, Calif: SAGE Publications.
- Newsom, J. T. (2015). Longitudinal Structural Equation Modeling: A Comprehensive Introduction. Routledge.
- Steele, F. (2008). Multilevel models for longitudinal data. Journal of the Royal Statistical Society A: Statistics in Sociey. 171:1, 5-19.
- Plewis, I. (2009) Statistical modelling for structured longitudinal designs. In Lynn, P. (ed.) (2009). Methodology of Longitudinal Surveys. Chichester: John Wiley. pp. 287-302.
- Plewis, I. (2010). Growth Modeling. In: Peterson, P. et al. (eds.) (2010). International Encyclopedia of Education (3rd. ed.). Elsevier. pp. 203-209.