SIB Swiss Institute of Bioinformatics

Mixed-effects modeling and longitudinal data analysis

Rachel Marcone

Translational Data Science group, SIB Swiss Institute of Bioinformatics Aug, 2022 - Lausanne

Repeated Measures and Longitudinal Data

- In repeated measures designs, measurements are taken repeatedly on each individual
- When these measurements are taken over time, it is called *longitudinal* design
- Typically various covariates concerning the individuals are recorded
  - 3 types of predictors
    - Time (or some notion thereof)
    - Time-variant
    - Time-invariant
- A general goal is to determine how a response depends on the covariates <u>over time</u>

Studying Change over Time

- Study can be experimental or observational
- Data can be collected prospectively or retrospectively
- Time can be measured in a variety of units (e.g. weeks, months, years, semesters, sessions, etc.)
- Nonetheless, need three or more waves of data
  - <u>More waves is always better</u>. If your data has only three waves, you must fit simpler models with stricter assumptions (e.g. linear). Additional waves allow for more flexible models with less restrictive assumptions.
  - <u>There is nothing sacrosanct about equal spacing</u>. If you expect rapid nonlinear change during some time periods, you should collect more data at those times.
  - <u>The resultant data need not be balanced</u>. In other words, each person need not have the same number of waves.

Traditional Methods have Limitations

- Traditional techniques for longitudinal data analysis, such as repeated measures ANOVA, can handle only complete data cases
- The assumptions of repeated measures ANOVA, in particular the assumption of sphericity/circularity, are often too restrictive for longitudinal data
  - Circularity refers to the condition where the variances of the differences between all pairs of within-subject conditions are equal
  - Departure from circularity results in inflated F-ratios
- Far less flexible in handling complex data structures compared to multi-level / mixed-effects modeling

Sphericity explained

Figure 1

Patient	Tx A	Tx B	Tx C	Tx A – Tx B	Tx A – Tx C	Tx B – Tx C
1	30	27	20	3	10	7
2	35	30	28	5	7	2
3	25	30	20	–5	5	10
4	15	15	12	0	3	3
5	9	12	7	–3	2	5
			Variance:	17	10.3	10.3

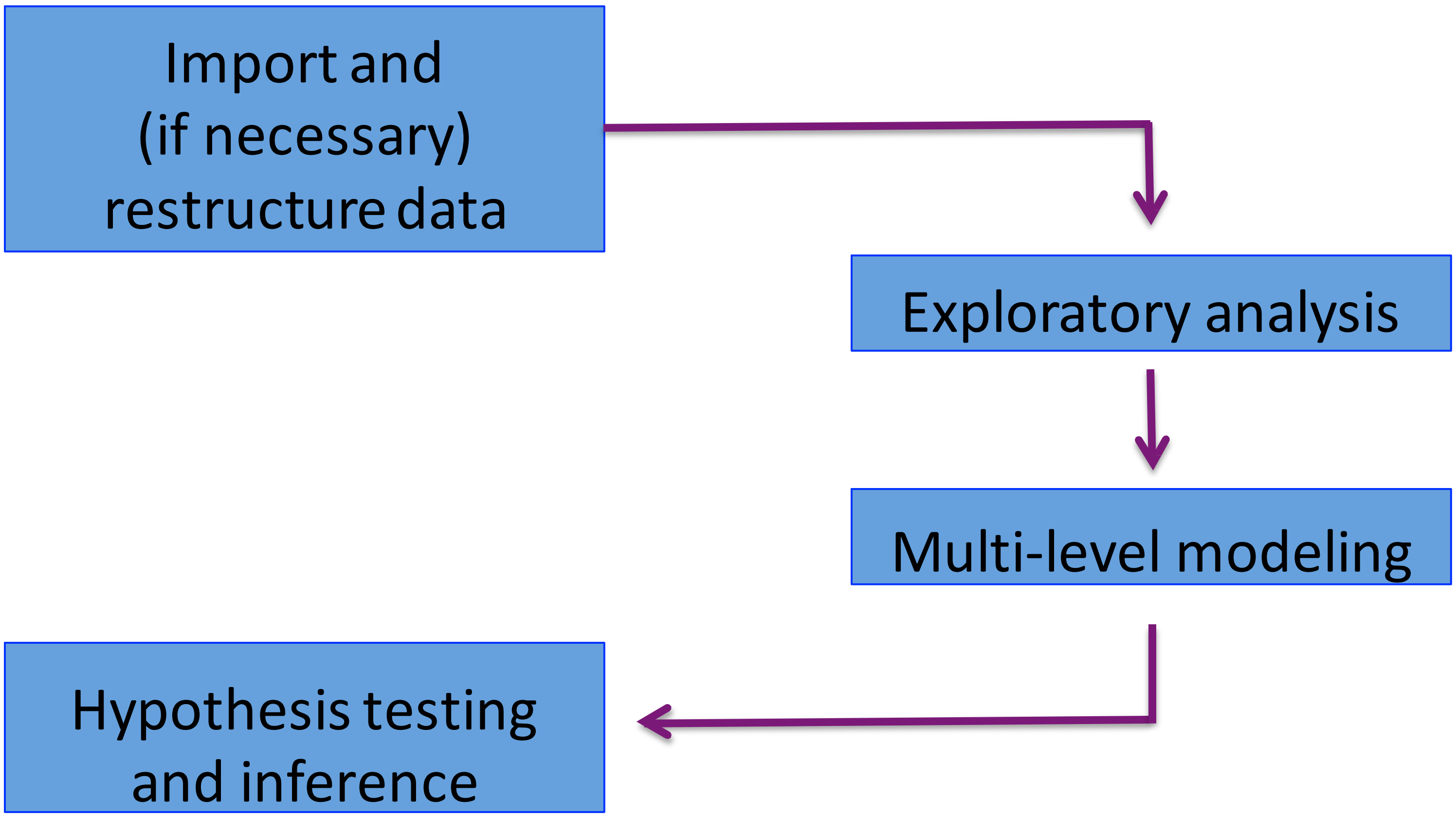
Multi-level / Mixed-effects Modeling of Longitudinal Data

- The mixed model approach to analyzing longitudinal data was commenced with the paper of Laird and Ware (1982)
- Most of the work that has been undertaken to model longitudinal data has been <u>parametric</u>, in the sense that the effects of continuous covariates have been modeled <u>linearly</u> or by using some <u>parametric nonlinear</u> model
- An alternative to nonlinear mixed modeling is to incorporate smoothing methods (aka semi-parametric mixed models) → We'll come back to this after learning about GAMs

Example #1: Tolerance data set

- National Youth Survey (NYS; Raudenbush & Chan 1992)
- Participants filled a survey at ages 11, 12, 13, 14, and 15 to provide their tolerance of deviant behavior
- Using a 4-point scale, where 1 = very wrong, 2 = wrong, 3 = a little bit wrong, and 4 = not wrong at all, participants indicated whether it was wrong for someone their age to
  - i. Cheat on tests
  - ii. Purposely destroy property of others
  - iii. Use marijuana
  - iv. Steal something
  - etc.
- Response (i.e. tolerance) was computed as respondent's average score
- Additional covariates
  - Gender (1=male and 0=female)
  - Exposure, representing respondent's self-reported exposure to deviant behavior at age 11, also on a 0-4 scale

Typical Workflow



"Tidy datasets are all alike but every untidy dataset is untidy in its own way." – Hadley Wickham

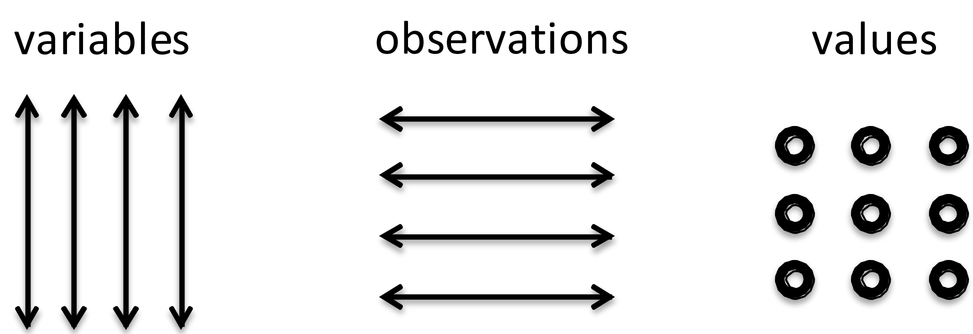
Wide format

	id	tol11	tol12	tol13	tol14	tol15	male	exposure
1	9	2.23	1.79	1.90	2.12	2.66	0	1.54
2	45	1.12	1.45	1.45	1.45	1.99	1	1.16
3	268	1.45	1.34	1.99	1.79	1.34	1	0.90
4	314	1.22	1.22	1.55	1.12	1.12	0	0.81
5	442	1.45	1.99	1.45	1.67	1.90	0	1.13
6	514	1.34	1.67	2.23	2.12	2.44	1	0.90
7	569	1.79	1.90	1.90	1.99	1.99	0	1.99
8	624	1.12	1.12	1.22	1.12	1.22	1	0.98
9	723	1.22	1.34	1.12	1.00	1.12	0	0.81
10	918	1.00	1.00	1.22	1.99	1.22	0	1.21
11	949	1.99	1.55	1.12	1.45	1.55	1	0.93
12	978	1.22	1.34	2.12	3.46	3.32	1	1.59
13	1105	1.34	1.90	1.99	1.90	2.12	1	1.38
14	1542	1.22	1.22	1.99	1.79	2.12	0	1.44
15	1552	1.00	1.12	2.23	1.55	1.55	0	1.04
16	1653	1.11	1.11	1.34	1.55	2.12	0	1.25

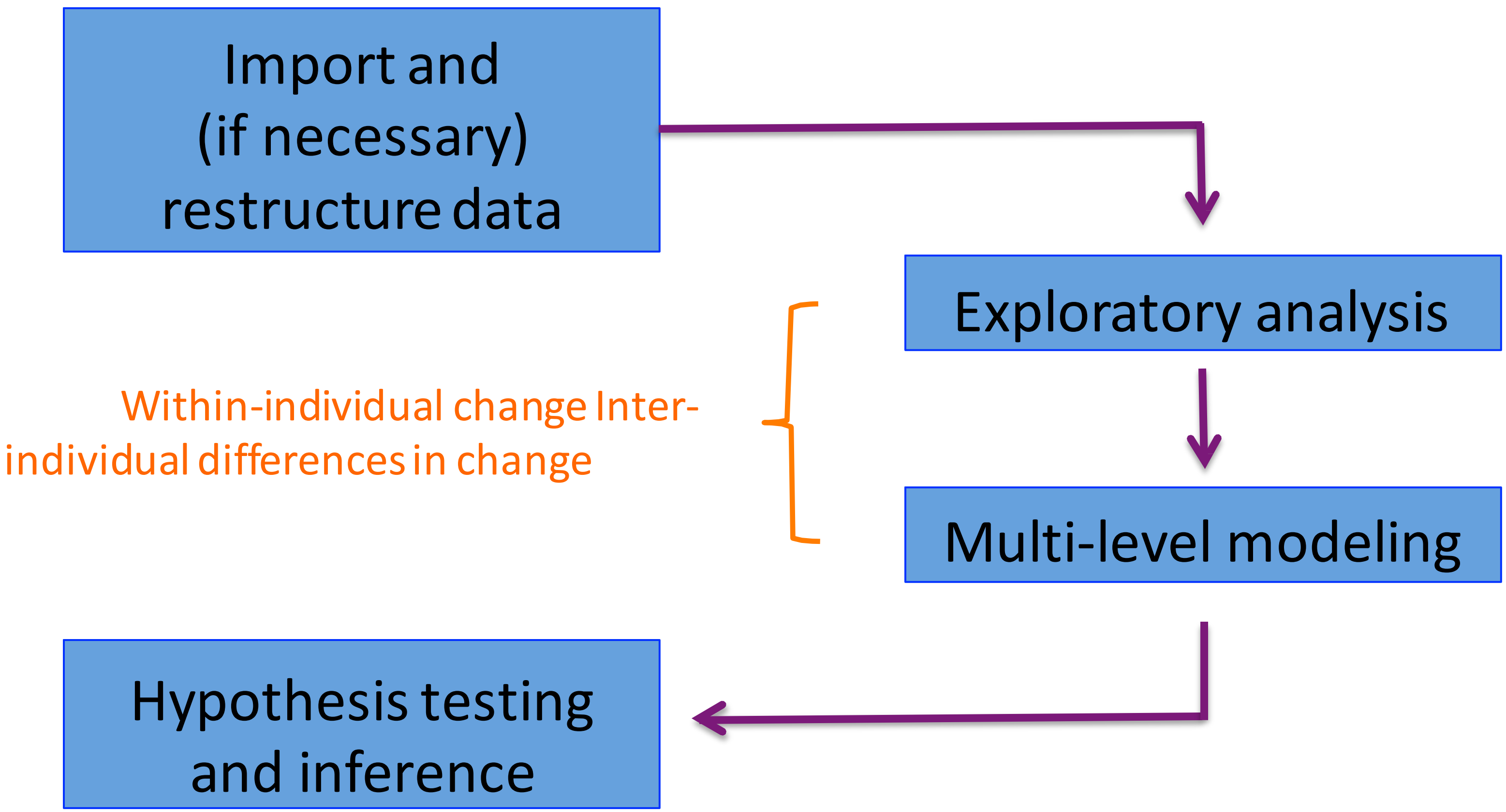
tidying →

Long format

	id	age	tolerance	male	exposure	time
1	9	11	2.23	0	1.54	0
2	9	12	1.79	0	1.54	1
3	9	13	1.90	0	1.54	2
4	9	14	2.12	0	1.54	3
5	9	15	2.66	0	1.54	4
6	45	11	1.12	1	1.16	0
7	45	12	1.45	1	1.16	1
8	45	13	1.45	1	1.16	2
9	45	14	1.45	1	1.16	3
10	45	15	1.99	1	1.16	4
11	268	11	1.45	1	0.90	0
12	268	12	1.34	1	0.90	1
13	268	13	1.99	1	0.90	2
14	268	14	1.79	1	0.90	3
15	268	15	1.34	1	0.90	4
16	314	11	1.22	0	0.81	0



Typical Workflow



A Framework for Analyzing Longitudinal Data

- Despite the unique set of outcome, predictors, and research questions, each longitudinal study poses an identical pair of questions:
  - **Q1) Within-individual change**
    - How does the outcome change over time?
    - E.g. how is each individual's pattern of tolerance over time? Is individual change linear or non-linear? Is it consistent over time or does it fluctuate?
  - **Q2) Inter-individual differences in change**
    - Can we predict differences in these changes?
    - E.g. do boys and girls experience different patterns of tolerance? Does baseline exposure affect boys and girls in the same way?

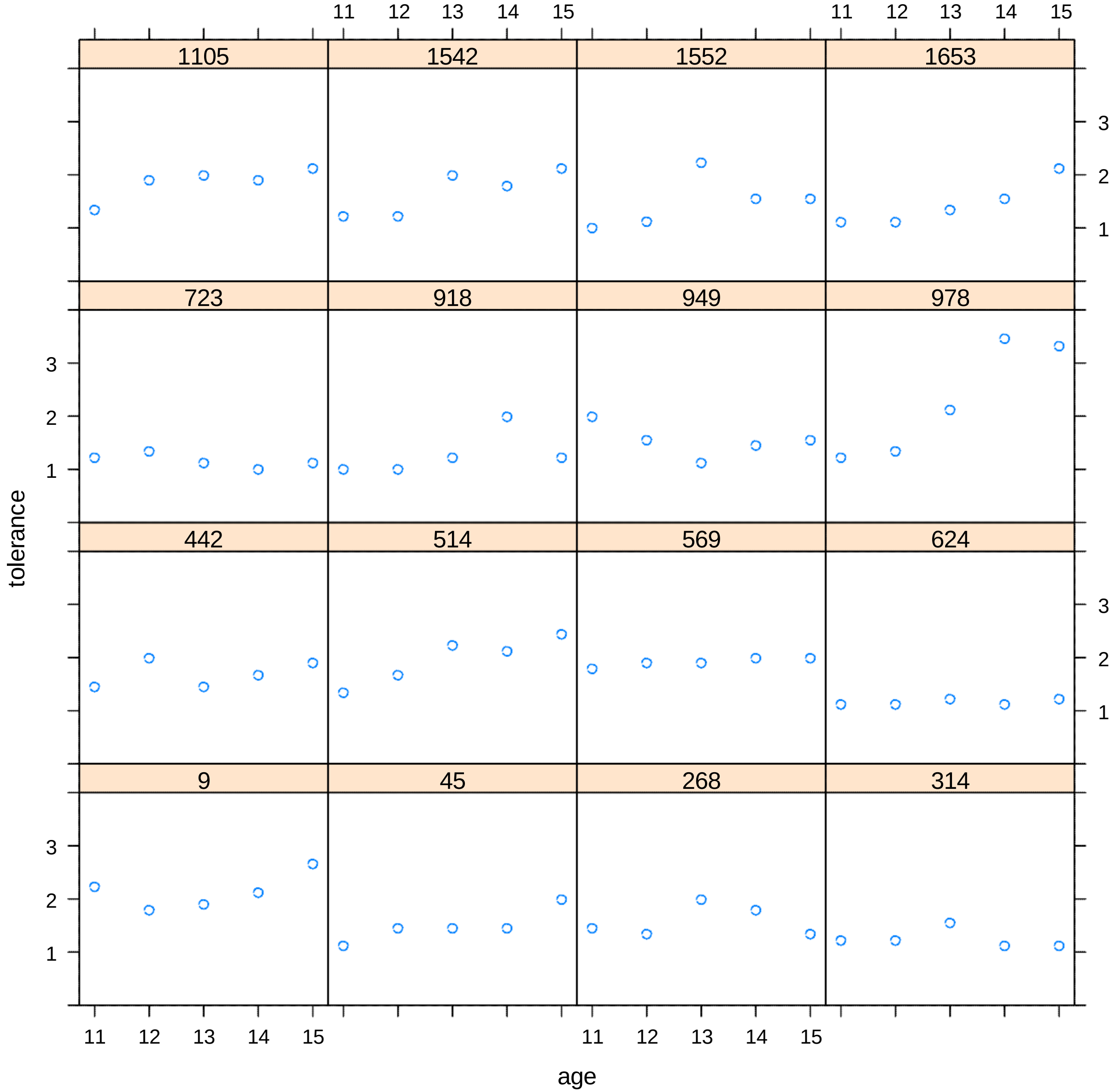
A Framework for Analyzing Longitudinal Data

- Despite the unique set of outcome, predictors, and research questions, each longitudinal study poses an identical pair of questions:
  - **Q1) Within-individual change**
    - How does the outcome change over time?
    - E.g. how is each individual's pattern of tolerance over time? Is individual change linear or non-linear? Is it consistent over time or does it fluctuate?
  - **Q2) Inter-individual differences in change**
    - Can we predict differences in these changes?
    - E.g. do boys and girls experience different patterns of tolerance? Does baseline exposure affect boys and girls in the same way?

Descriptive Analysis of Individual Change over Time

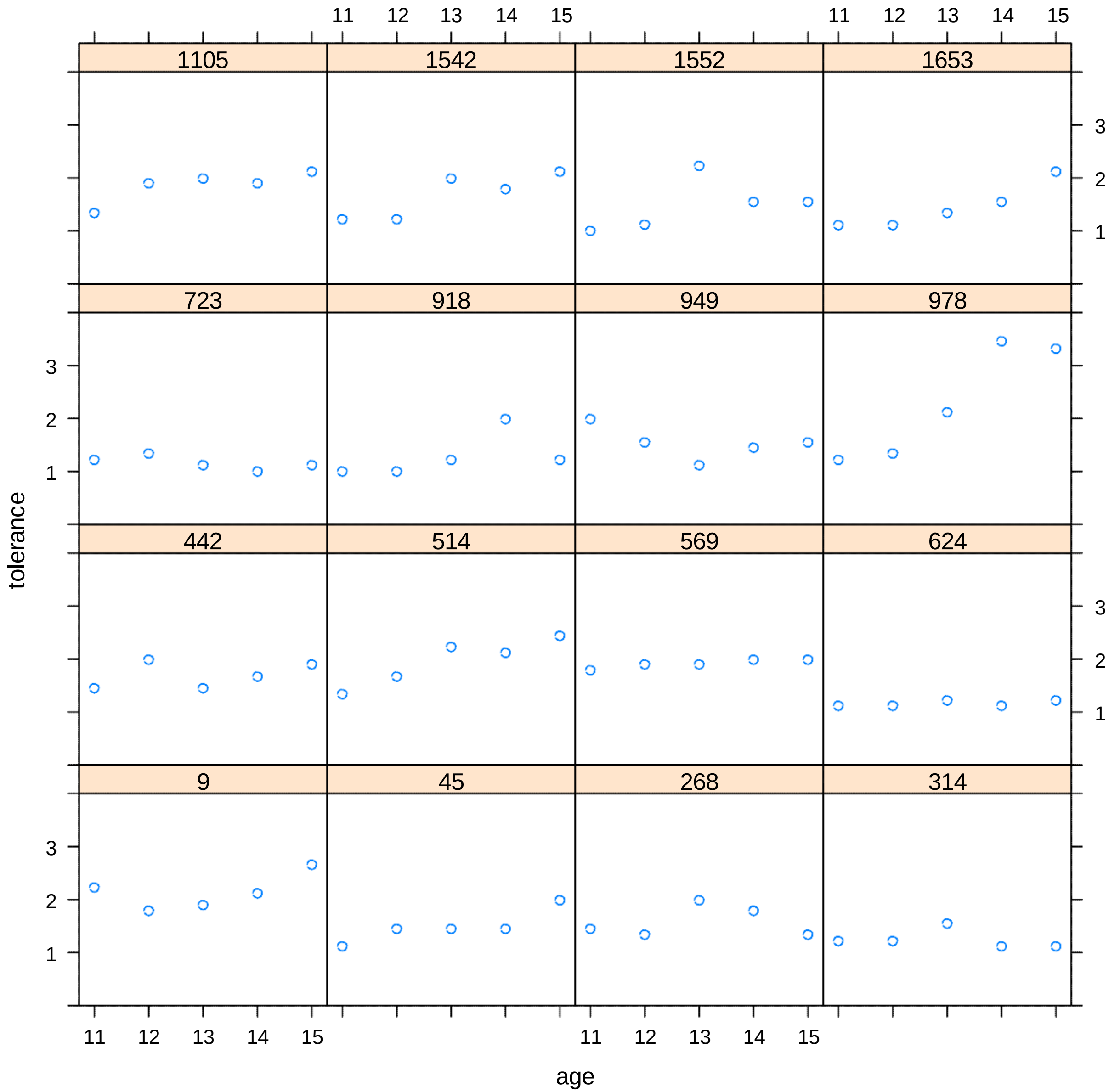
- **STEP1: Start with scatterplots (response vs. time) $\boxed{\text{at the individual level}}$** **Q1**
  - Easier to discern if sets of individuals are plotted on the same panel
  - Better to use identical axes across individuals
  - In large data sets, may have to inspect a random subset of individuals

Descriptive Analysis of Individual Change over Time



Descriptive Analysis of Individual Change over Time

On average the tolerance appears to increase with time



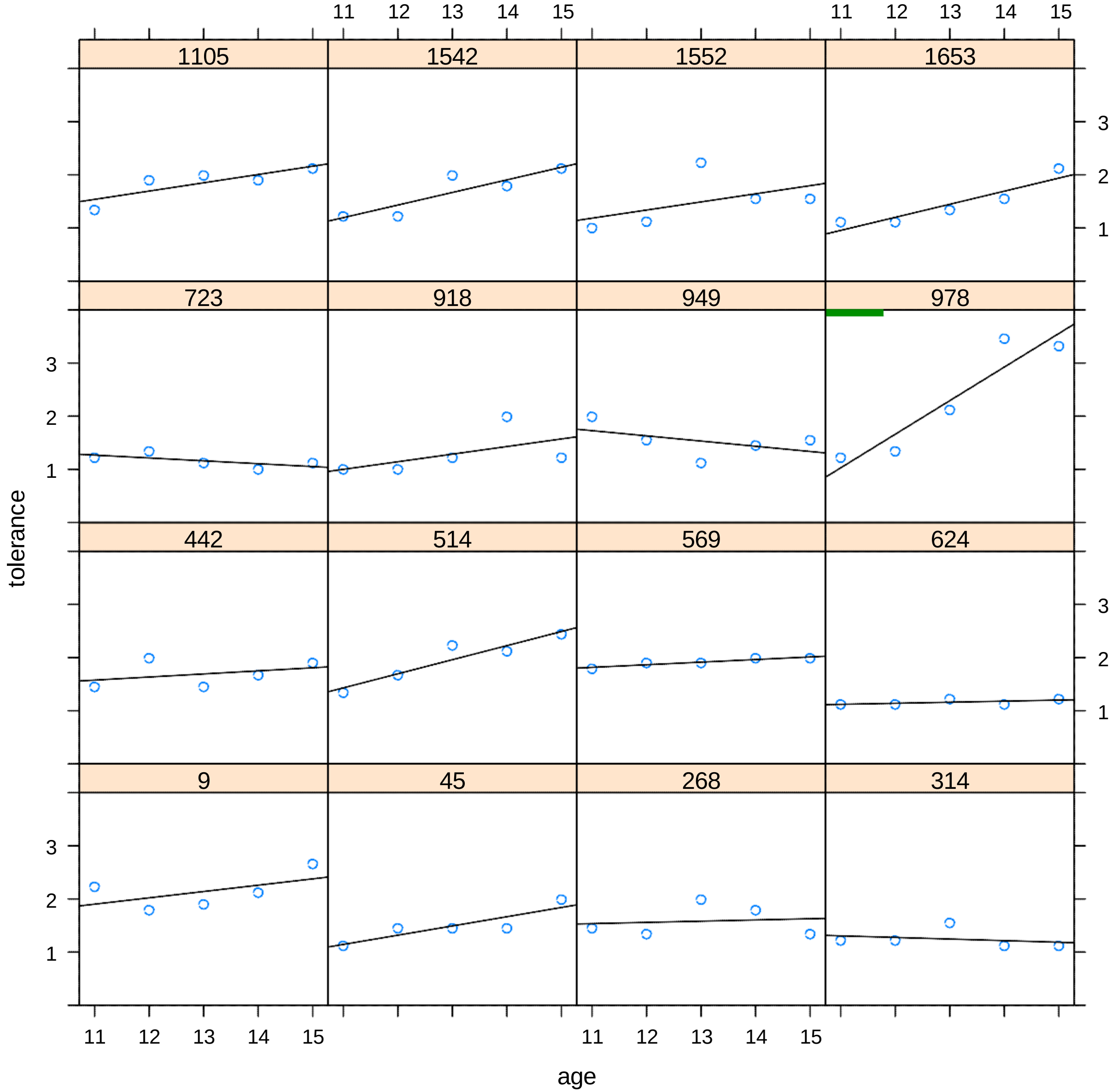
Except for #978, tolerance generally stays in the lower portion

Descriptive Analysis of Individual Change over Time

- **STEP2: Use a trajectory to summarize each person's temporal pattern of change**
  - Nonparametric approach (e.g. smoothing spline)
    - Requires no assumptions
    - Letting the data speak for themselves
  - Parametric approach (e.g. linear, quadratic, etc.)
    - Requires assumptions
    - But instead provides numeric summaries of the trajectories
- Exploratory analysis often suggests that different people require different functions
  - Measurement error makes it difficult to discern if individual patterns are true signal or simply due to random fluctuation
  - Of course, fitting person-specific models, one individual at a time, is not the most efficient use of longitudinal data; That's why we will use mixed-effects modeling eventually

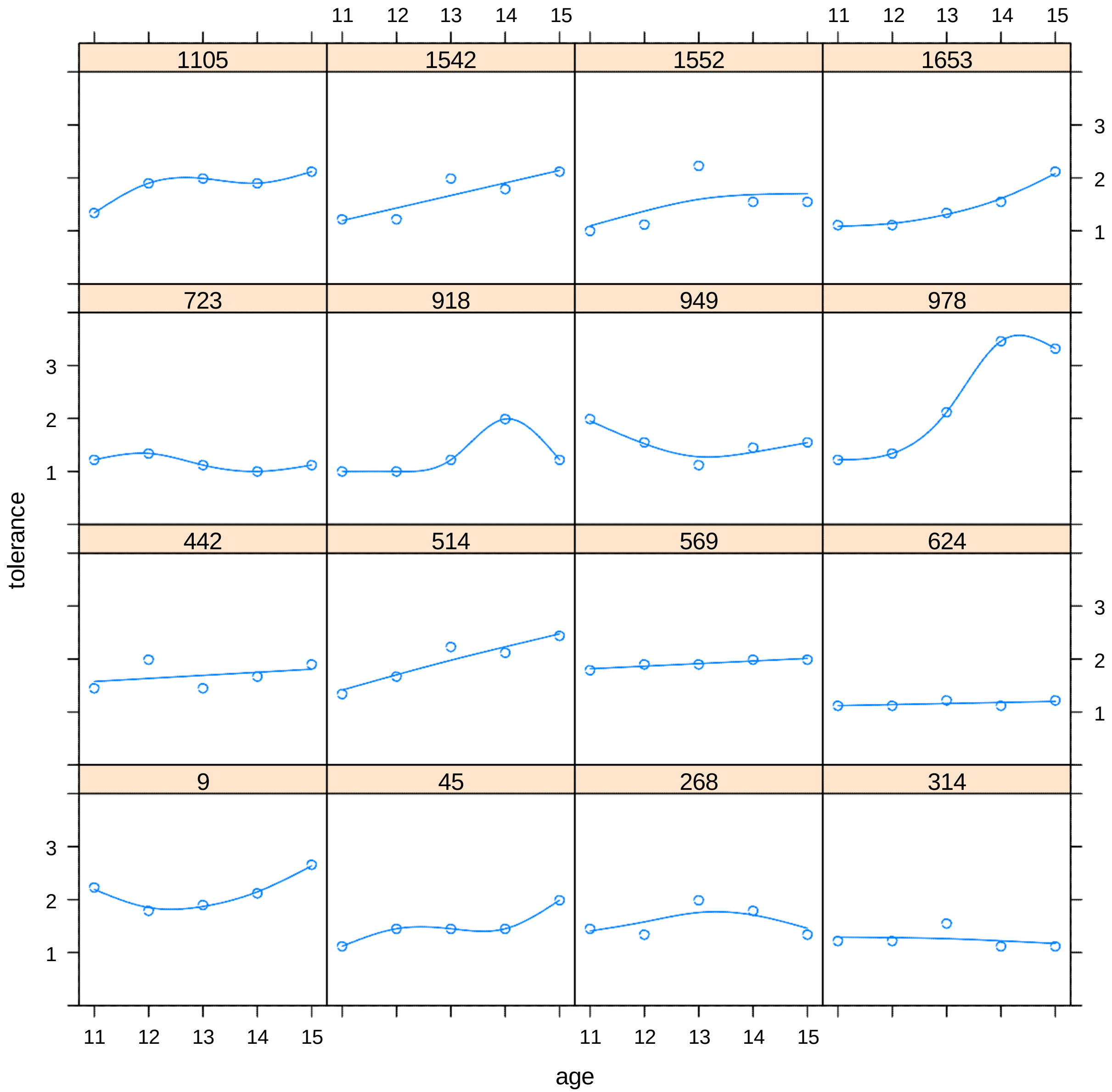
Descriptive Analysis of Individual Change over Time

linear fit



intercept = 1.03
slope = 0.63
res.var = 0.17
$R^2 = 0.89$

Descriptive Analysis of Individual Change over Time



smoothing
spline

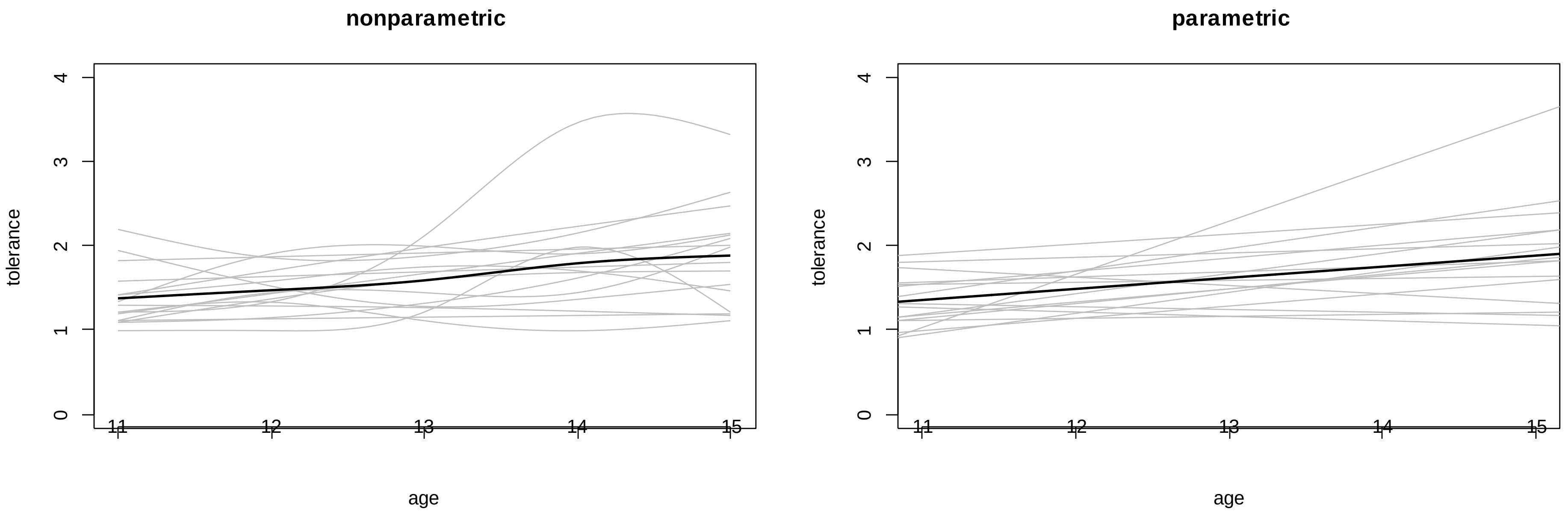
A Framework for Analyzing Longitudinal Data

- Despite the unique set of outcome, predictors, and research questions, each longitudinal study poses an identical pair of questions:
  - **Q1) Within-individual change**
    - How does the outcome change over time?
    - E.g. how is each individual's pattern of tolerance over time? Is individual change linear or non-linear? Is it consistent over time or does it fluctuate?
  - **Q2) Inter-individual differences in change**
    - Can we predict differences in these changes?
    - E.g. do boys and girls experience different patterns of tolerance? Does baseline exposure affect boys and girls in the same way?

Exploring Differences in Change across People

- Does everyone change in the same way, or do the trajectories of change differ substantially across people? **Q2**
- **STEP3: Plot, on a single graph, the entire set of individual trajectories, and compute an average change trajectory**
  i. Discretize time on a reasonably refined grid
  ii. Estimate individual trajectories on the grid
  iii. Average individual estimates for each point on the grid
  iv. Apply the same smoothing algorithm, nonparametric or parametric, used to obtain individual trajectories
- NEVER infer the shape of the individual change trajectories from the shape of their averages
  – The only kind of trajectories for which the "average of the curves" is identical to the "curve of the averages" is one whose mathematical representation is *linear in the parameters* [Keats 1983]
  – We examine the averages simply for comparison, not to learn anything about the underlying shapes of the individual trajectories

Exploring Differences in Change across People



Exploring Differences in Change across People

- **STEP4: Use the results of exploratory analysis to (re-)frame questions about change**
  - Adopting a parametric model for individual change allows us to express generic questions about inter-individual differences in change as specific questions about the parameters of the model
  - E.g. in the case of linear fit, it is often helpful to examine
    1. Sample means of the estimated intercepts and slopes
    2. Sample variances (or SD) of the estimated intercepts and slopes
    3. Sample correlation between the estimated intercepts and slopes

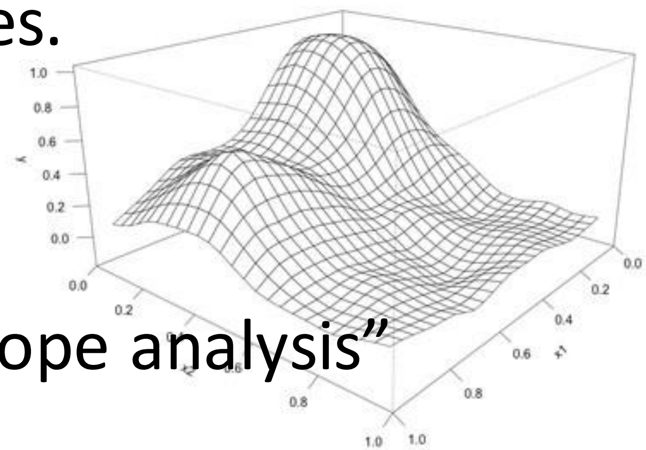
	intercept	slope
mean	1.36	0.13
SD	0.30	0.17
Bivariate corr		-0.45

Exploring Differences in Change across People

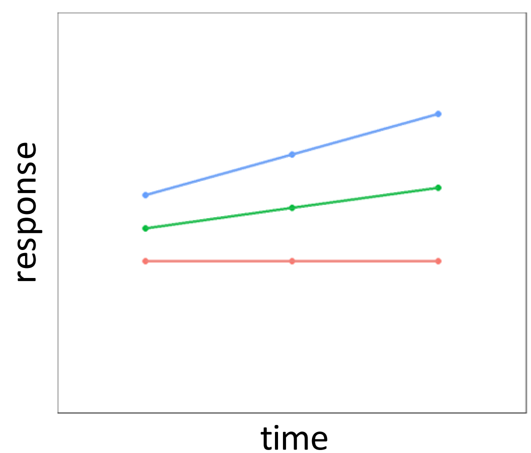
- **STEP5: Explore the relationship between change and <u>time-invariant</u> predictors**
  - Allows us to uncover systematic patterns in the individual change trajectories corresponding to inter-individual variation observed in personal trajectories, e.g.
    - Examining differences by gender allows us to assess
      - Whether boys or girls are initially more tolerant of deviant behavior
      - Whether boys and girls tend to have different annual rates of change
    - Examining differences by baseline exposure allows us to assess
      - Whether a child' initial level of tolerance is associated with baseline exposure
      - Whether a child' rate of change in tolerance is associated with baseline exposure
  - If a predictor is continuous you can temporarily categorize it

Exploring Differences in Change across People

- What about <u>time-variant</u> predictors?
  - For example via loess (with multiple predictors) or thin plate splines (from mgcv). See persp() in base R graphics for plotting 3D surfaces.



  - Another approach is “simple slope analysis”



1SD above mean
mean
1SD below mean

Time-variant
predictor

LOESS

Stands for **locally estimated scatterplot smoothing**
Non parametric
Combines multiple regression models in a K-nearest neighbor based meta-model

Thin Plate splines

Thin plate splines (TPS) are a technique for data interpolation and smoothing.

Advantages :

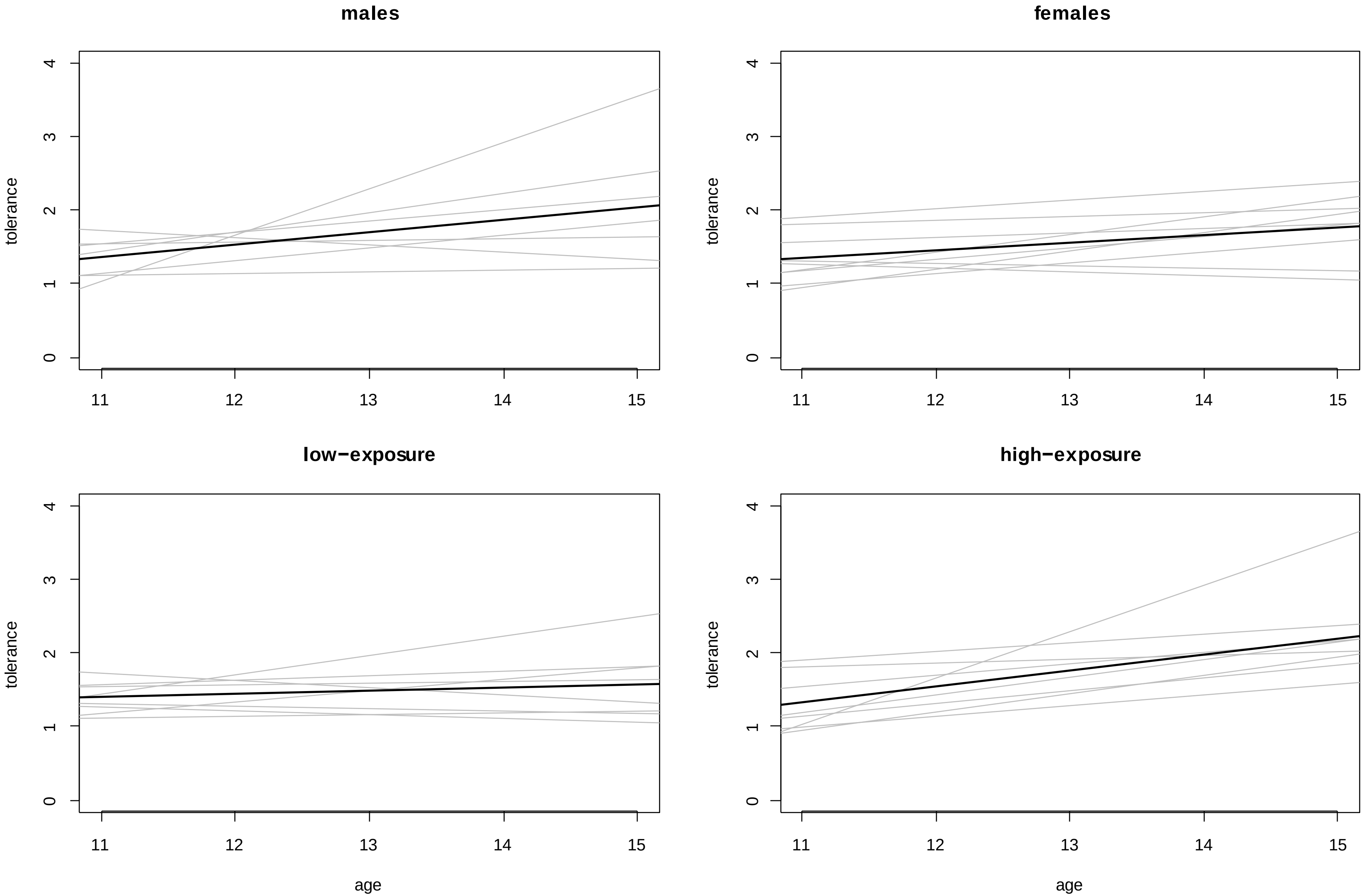
It produces smooth surfaces, which are infinitely differentiable.

There are no free parameters that need manual tuning.

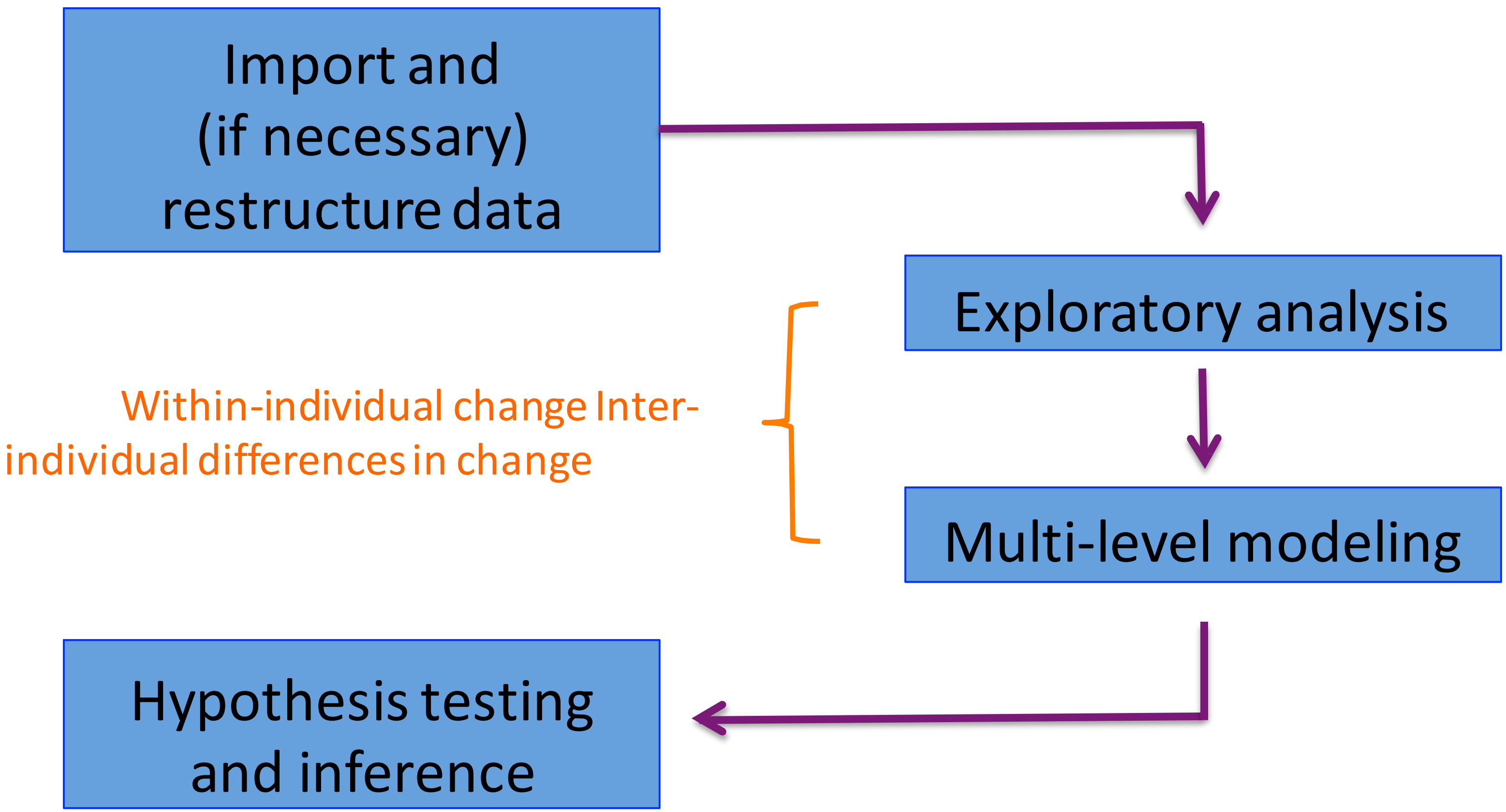
It has closed-form solutions for both warping and parameter estimation.

There is a physical explanation for its energy function.

Exploring Differences in Change across People



Typical Workflow



Exercises

Tolerance.R -> Open and Follow the instruction

Corrections in a few…

Mixed Models

Mixed-effects models provide a flexible and powerful tool for the analysis of grouped data, including:

- blocked designs
- repeated measures (each subject measured for each condition; individuals are `blocks')
- Longitudinal data (measures repeated over time)
- multilevel data

Mixed Models

Offer flexibility in modeling within-group correlation often present in grouped data

- Handle balanced and unbalanced data in a unified
framework
- There is reliable, efficient software for fitting

They are characterized by statistical models containing fixed and random effects.

Usually represented by the following formula

$$\boldsymbol{y} = X\boldsymbol{\beta} + Z\boldsymbol{u} + \boldsymbol{\epsilon}$$

Mixed Modeling: ML vs. REML

- Choice of methods for parameter estimation/prediction
  - Maximum Likelihood (ML)
  - Restricted Maximum Likelihood (REML)
- The general wisdom is that
    - ML produces more accurate estimates of the fixed parameters, whereas REML produces more accurate predictions of random variances
    - Therefore, the choice depends on whether the hypotheses are focused on the fixed parameters or random variances
    - In practice, the choice of ML or REML will make only small differences to the parameters
    - **If you want to compare models, you <u>MUST</u> use ML**

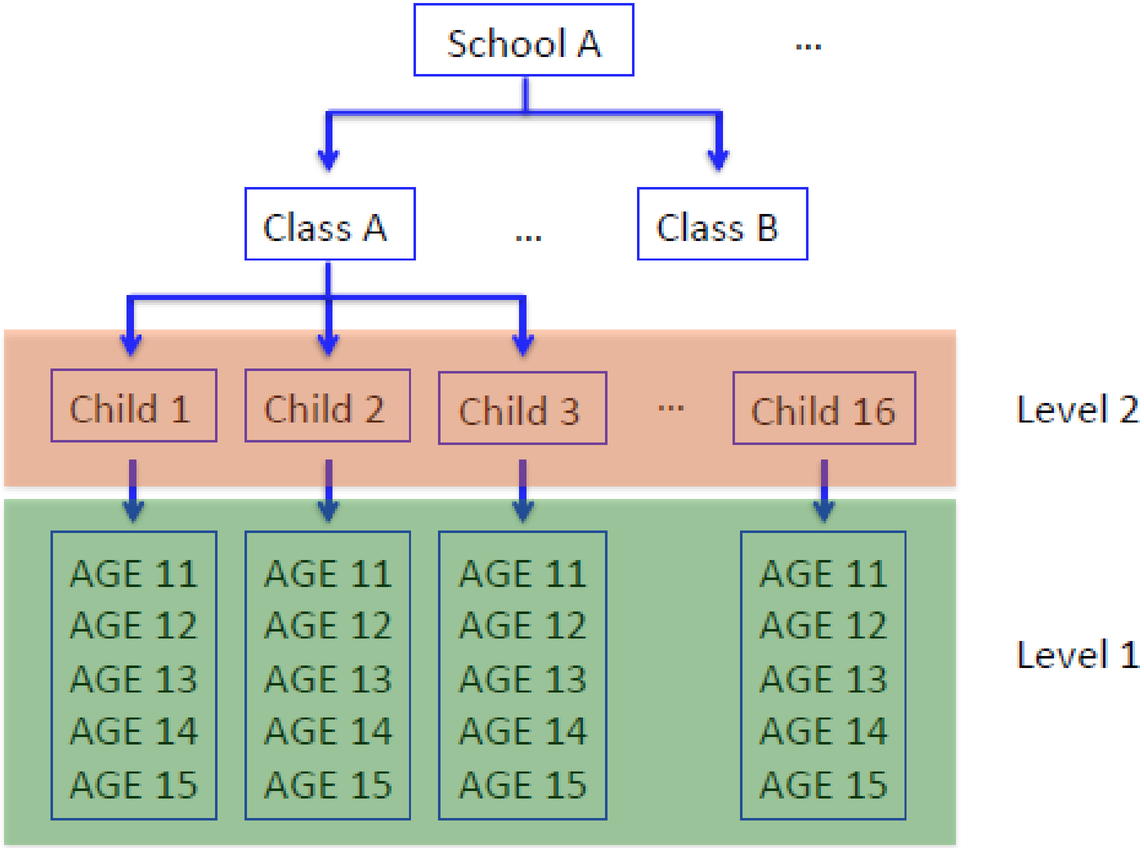
Mixed Modeling: Model Selection

- Assessing the fit and comparing multi-level models
  - Likelihood ratio test based on ML theory
    - R reports the log-likelihood (LL)
    - Essentially the smaller the value of LL the better
  - R also produces two adjusted LL values, which can be interpreted in a similar manner
    - AIC: Akaike's Information Criterion
      - A goodness of fit measure corrected for model complexity
    - BIC: Bayesian Information Criterion
      - Comparable to AIC but slightly more conservative (i.e. it corrects more harshly for the number of parameters)
    - Neither AIC or BIC are intrinsically interpretable; they are only useful in comparing models. Smaller values in both cases mean a better-fitting model
    - Recommended approach: start simple!

Multi-level Modeling Approach to Longitudinal Data

- A key assumption of standard linear regression is the assumption of independently distributed error terms for the individual observations within a sample
  - Essentially means that there are no relationships among individual observations for the dependent variable once the independent variables are accounted for
- In real world, data often has a hierarchical structure, hence the name “multi-level”
  - It simply means that some variables are *clustered* or *nested* within other variables
  - In longitudinal design, this hierarchy (in part) stems from repeated measures obtained from the same individual over time

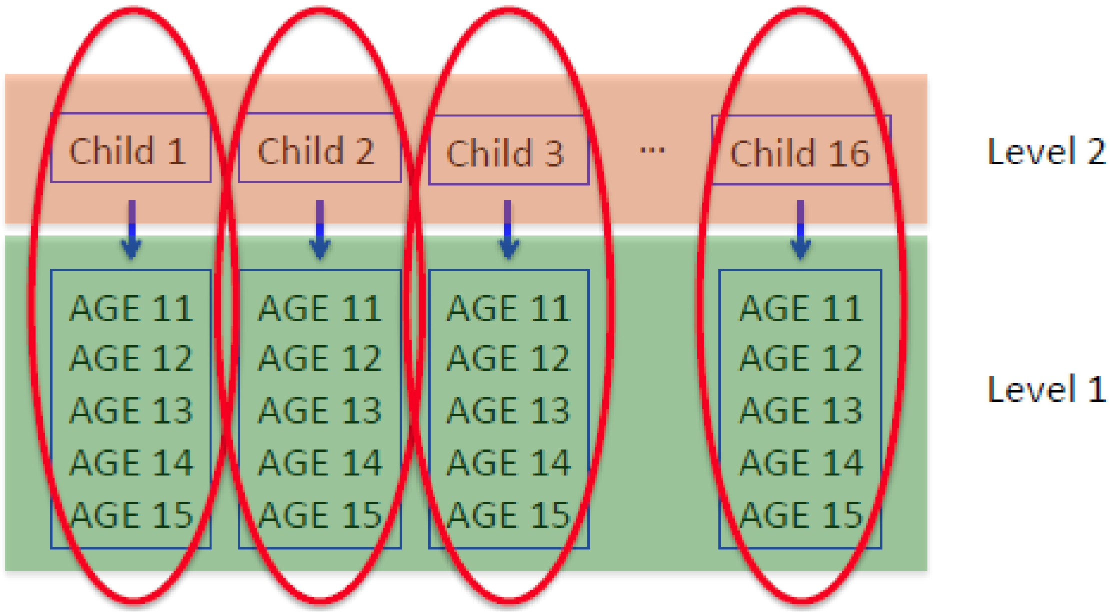
Multi-level Modeling Approach to Longitudinal Data



Multi-level Modeling Approach to Longitudinal Data

- When dealing with multi-level data, the assumption of independent errors is violated
  - i.e. the potential inter-individual correlation (correlation among repeated measures obtained from the same individual) may result in inappropriate estimate of model parameters
- Moreover, by ignoring multi-level structure of data we may miss important relationships involving each level in the data
- **Solution: When data has multi-level structure, we allow the parameters of the model to vary between clusters**
  - Fixed vs. random coefficients (mixed-effects modeling)
  - Packages for multi-level modeling in R
    - nlme (allows for flexible modeling of the covariance structure)
    - lme4

Multi-level Modeling Approach to Longitudinal Data



Allowing the parameters of the model to vary between clusters:
Either allowing the intercepts to vary among children,
or allowing the slopes to vary among children,
or allowing both intercepts and slopes to vary among children.

Instead of estimating 16 intercepts and 16 slopes, we will estimate one fixed intercept, one fixed slope, one random intercept and one random slope.

Multi-level Modeling Approach to Longitudinal Data

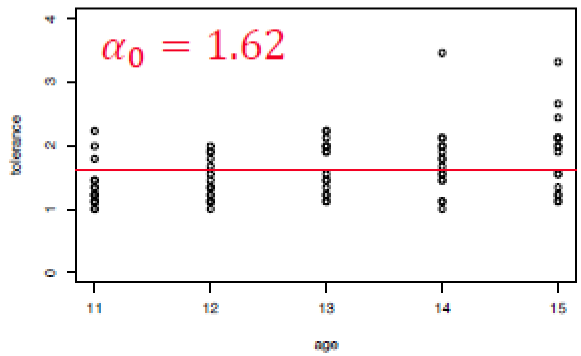
$$y_{ij} = \boxed{\alpha} + \varepsilon_{ij} \quad for\ individual\ i\ at\ time\ j$$

$$y_{ij} = \alpha_0 + \varepsilon_{ij} \qquad \text{(Model 00)}$$

α_0 *is a fixed variable*

$$y_{ij} = (\alpha_0 + \alpha_1) + \varepsilon_{ij} \qquad \text{(Model 01)}$$

α_0 *is a fixed variable whereas* α_1 *is a random variable*



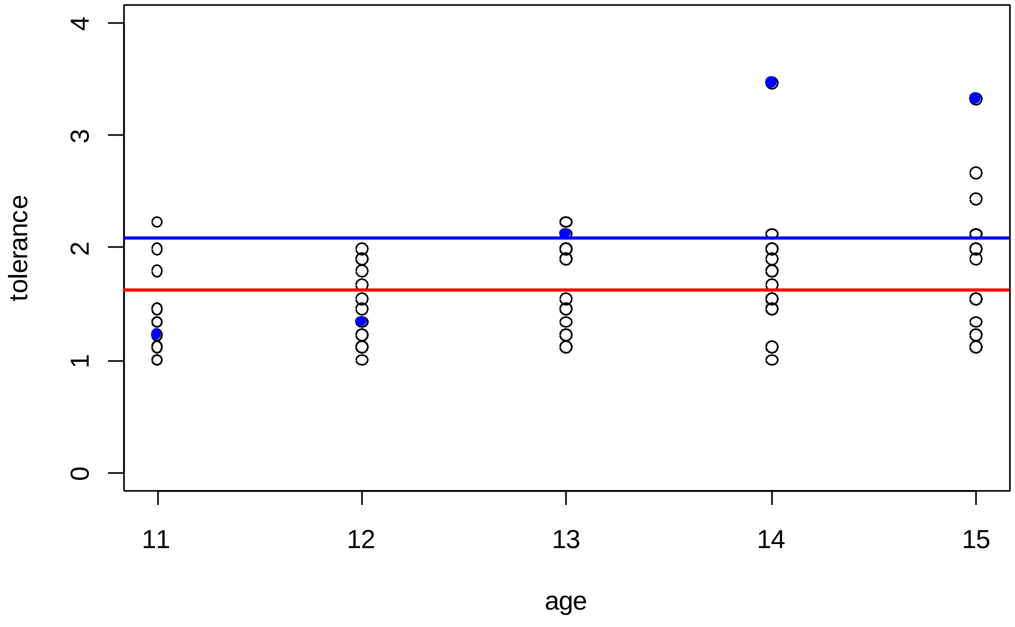

```
> anova(fit.00,fit.01)
       Model df      AIC      BIC    logLik   Test  L.Ratio p-value
fit.00     1  2 117.7198 122.4839 -56.85992
fit.01     2  3 109.0219 116.1679 -51.51093 1 vs 2 10.698  0.0011
```

Multi-level Modeling Approach to Longitudinal Data

$$y_{ij} = \alpha + \varepsilon_{ij} \quad for\ individual\ i\ at\ time\ j$$

$$y_{ij} = (\alpha_0 + \alpha_1) + \varepsilon_{ij}$$ (Model 01)

$$\alpha_0\ is\ a\ fixed\ variable\ whereas\ \alpha_1\ is\ a\ random\ variable$$



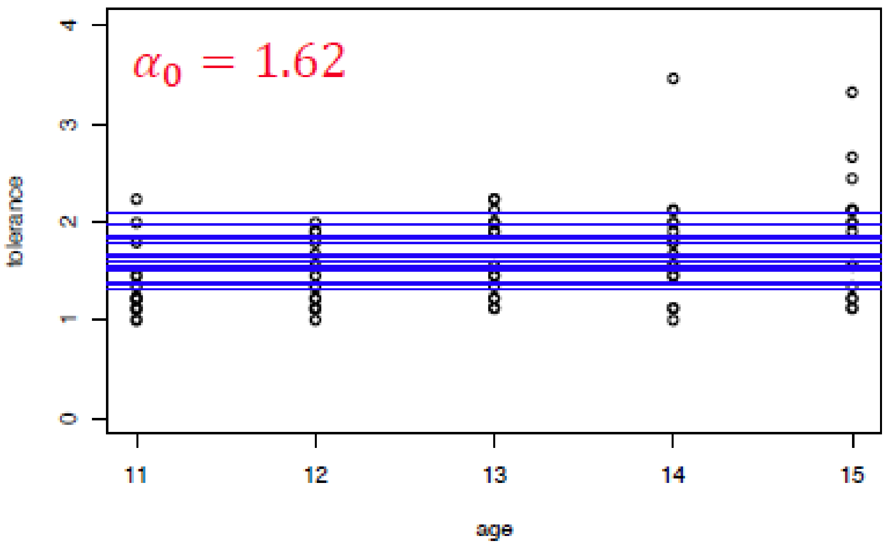
Individual #978

Multi-level Modeling Approach to Longitudinal Data

$$y_{ij} = \alpha + \varepsilon_{ij} \quad for\ individual\ i\ at\ time\ j$$

$$y_{ij} = (\alpha_0 + \alpha_1) + \varepsilon_{ij}$$ (Model 01)

α_0 *is a fixed variable whereas* α_1 *is a random variable*



sixteen individuals

Multi-level Modeling Approach to Longitudinal Data

$$y_{ij} = \alpha + \boxed{\beta} t_j + \varepsilon_{ij} \quad for\ individual\ i\ at\ time\ j$$

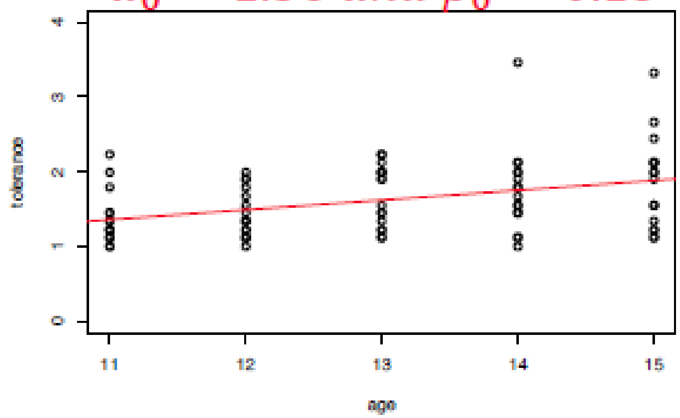
$$y_{ij} = \alpha_0 + \beta_0 t + \varepsilon_{ij} \qquad \text{(Model 03)}$$

$$\alpha_0\ and\ \beta_0\ are\ fixed\ variables$$

$$y_{ij} = \alpha_0 + (\beta_0 + \beta_1) t_j + \varepsilon_{ij} \qquad \text{(Model 04)}$$

$$\alpha_0\ and\ \beta_0\ are\ fixed\ variables\ whereas\ \beta_1\ is\ a\ random\ variable$$

$$\alpha_0 = 1.36\ and\ \beta_0 = 0.13$$



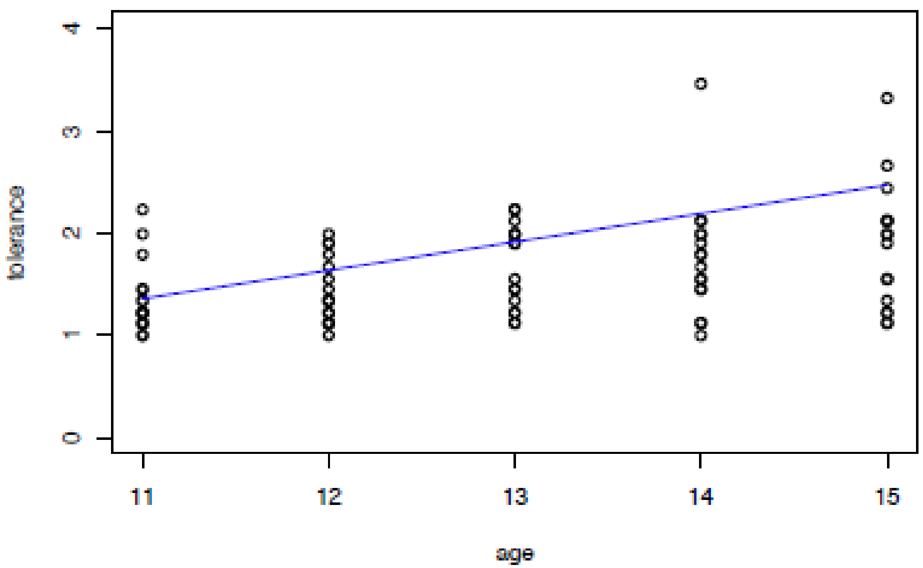

```
> anova(fit.03,fit.04)
       Model df       AIC       BIC    logLik   Test  L.Ratio p-value
fit.03     1  3 107.55383 114.69991 -50.77692
fit.04     2  4  74.93266  84.46077 -33.46633 1 vs 2 34.62117  <.0001
```

Multi-level Modeling Approach to Longitudinal Data

$$y_{ij} = \alpha + \beta t_j + \varepsilon_{ij} \quad for\ individual\ i\ at\ time\ j$$

$$y_{ij} = \alpha_0 + (\beta_0 + \beta_1)t_j + \varepsilon_{ij}$$ (Model 04)

α_0 *and* β_0 *are fixed variables whereas* β_1 *is a random variable*



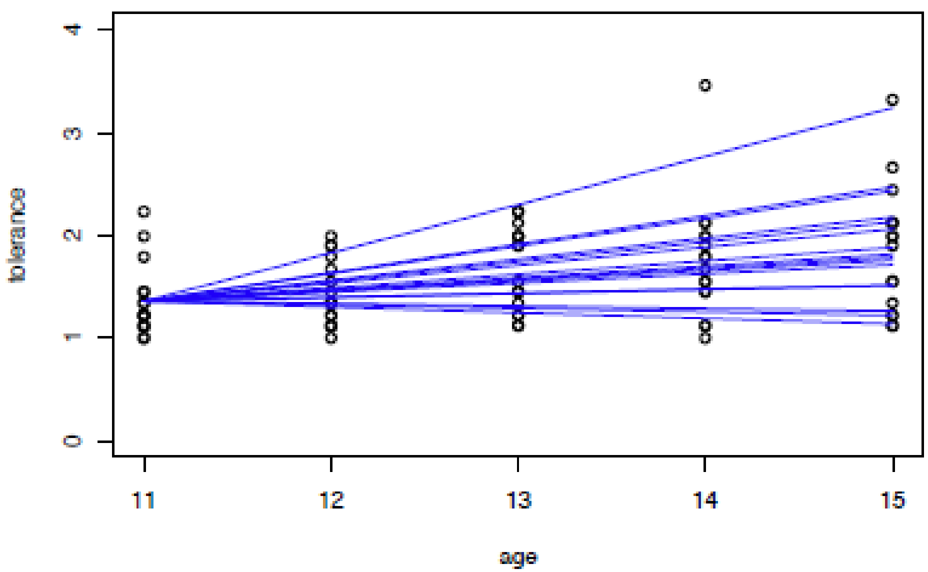
Individual #978

Multi-level Modeling Approach to Longitudinal Data

$$y_{ij} = \alpha + \beta t_j + \varepsilon_{ij} \quad for\ individual\ i\ at\ time\ j$$

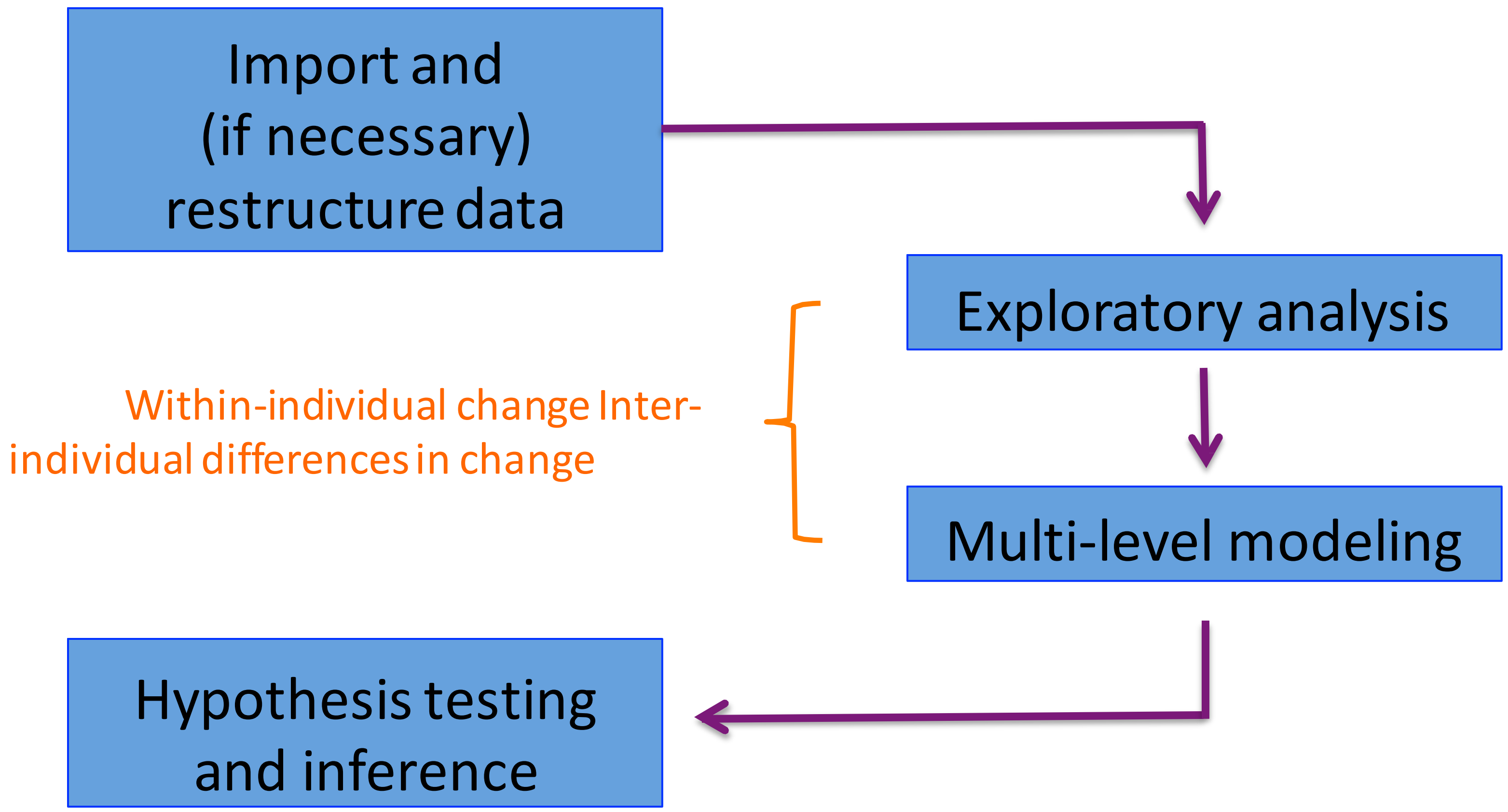
$$y_{ij} = \alpha_0 + (\beta_0 + \beta_1)t_j + \varepsilon_{ij}$$ (Model 04)

α_0 *and* β_0 *are fixed variables whereas* β_1 *is a random variable*



Sixteen individuals

Summary of a Typical Workflow



Multi-level Modeling Approach to Longitudinal Data

- Assumptions about the covariance structure of the data
  - Variance components
  - Diagonal
  - AR(1): typically used when data is measured over time
    - corAR1(…)
    - corCAR1(…)
    - corARMA(…)
  - Unstructured

We'll come back to this tomorrow!

Working Exercises in R

- #1: Corn Data
- #2: Tolerance2.R
- #3: Rat_Brain.R
- #4: “Beat the Blues” clinical trial
  Beat_the_Blues.R
- #5: Bone_mineral_density.R