Help! Statistics!

Introduction to Longitudinal Data Analysis

Sacha la Bastide-van Gemert
Medical Statistics and Decision Making
Epidemiology, UMCG

Help! Statistics! Lunch time lectures

What? Frequently used statistical methods and questions in a manageable timeframe for all researchers at the UMCG.
No knowledge of advanced statistics is required.

When? Lectures take place every 2nd Tuesday of the month, 12.00-13.00 hrs.

Who? Unit for Medical Statistics and Decision Making

When?	Where?	What?	Who?
Dec 12, 2017	Room 16	Propensity Scoring	C. zu Eulenburg
2018:			
Feb 13, 2018	.....	Regression to the mean and other pitfalls	H. Burgerhof
March 13, 2018	.....	.....	.....
…			

Slides can be downloaded from:
http://www.rug.nl/research/epidemiology/download-area

Introduction to longitudinal data analyses: overview

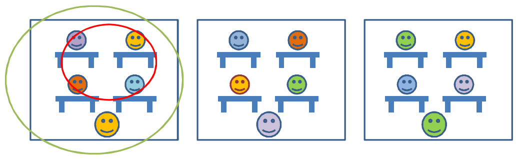
- What is longitudinal data?
- Why does it need a special approach?
  - revisiting the linear regression model
- Longitudinal data analysis: using summary measures
- Longitudinal data analysis: introduction of the multilevel model for change (mixed effects model)

What is longitudinal data? (1)

Clustered data

Clustered (or nested/multilevel/hierarchical/…) data:

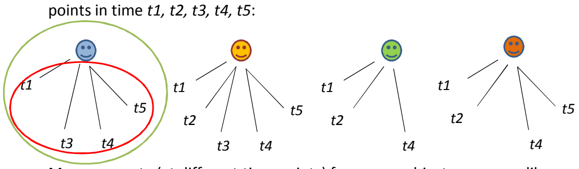
Example: several classrooms, within each classroom students



- (Results from) students from the same classroom are more alike than students from different classrooms: students are *nested* in classrooms
- Variables at student level: gender, SES, …
- Variables at classroom level: teacher effect, … -> multilevel data

What is longitudinal data? (2)

- Longitudinal data: several subjects, each measured at several (different) points in time *t1, t2, t3, t4, t5*:



- Measurements (at different time points) from one subject are more alike than measurements from different subjects: *measurements are nested within subjects*
- Variables at each time point: lengths, grades…
- Variables for each subject: gender, SES, … -> multilevel data

Longitudinal data: investigating change over time

- Change over time: natural (growth, ageing) or due to intervention (medication, diet, therapy)
- Longitudinal data: *outcome variable* consists of multiple measurements (the more, the better) of the same type at different time points
  Example: infants' lengths at *age_1, age_2, age_3,…*
- Additionally: independent *explanatory variables* or *covariates*
  Example: gender, treatment group, …

➢ *Key: investigating change requires longitudinal data (≠ cross-sectional data)*

Today: focus on continuous outcome variables

Example: adolescent alcohol use (Curran et al, 1997)*

- Sample of 82 adolescents:
  37 are children of an alcoholic parent (COAs), 45 are non-COAs
- Research design:
  - each child assessed 3 times (at ages 14, 15, and 16)
  - outcome: *alcuse* (continuous, ``alcohol use'' based on various items)
  - covariate (among others): *COA* (dichotomous)

- Research question:
  Do trajectories of adolescent alcohol use differ by parental alcoholism?

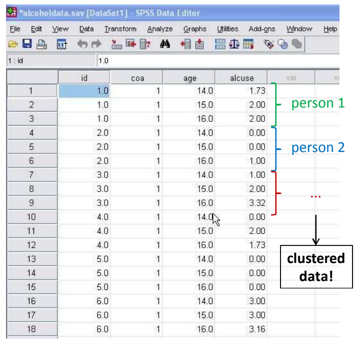
* Example from: Singer & Willet: Applied longitudinal data analysis. Modeling change and event occurence (Oxford, 2003)

Longitudinal data
The data-set: person-period format

The person-period format: for each person, each repeated measurement is stored as a new case

Here: 3 rows per person
- a time variable: *age*
- an outcome variable: *alcuse*
- a (time-independent) covariate: *coa*

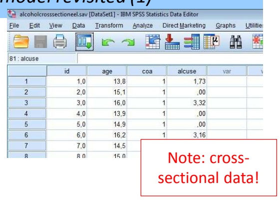
How to proceed?
Let's revisit (simple) linear regression analysis...

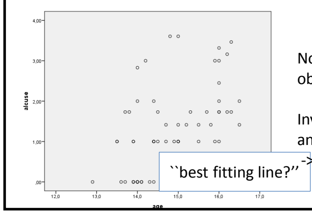


	id	coa	age	alcuse
1	1.0	1	14.0	1.73
2	1.0	1	15.0	2.00
3	1.0	1	16.0	2.00
4	2.0	1	14.0	0.00
5	2.0	1	15.0	0.00
6	2.0	1	16.0	1.00
7	3.0	1	14.0	1.00
8	3.0	1	15.0	2.00
9	3.0	1	16.0	3.32
10	4.0	1	14.0	0.00
11	4.0	1	15.0	2.00
12	4.0	1	16.0	1.73
13	5.0	1	14.0	0.00
14	5.0	1	15.0	0.00
15	5.0	1	16.0	0.00
16	6.0	1	14.0	3.00
17	6.0	1	15.0	3.00
18	6.0	1	16.0	3.16

Intermezzo
The linear regression model revisited (1)

A new data-set:
- a continuous outcome variable *Y* (here: *alcuse*)
- one or more explanatory variables *x1, x2, ...* (here: *age, COA*)





Now: for each adolescent *i* (= 1,...,82) one observation (*alcuse, age*) in the dataset

Investigating the relation between *age* and *alcuse*: a linear relationship?
-> scatterplot *age-alcuse*

Intermezzo
The linear regression model revisited (2)

Formally: we assume an underlying true population linear relationship, described by (subject *i*):

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2)$$

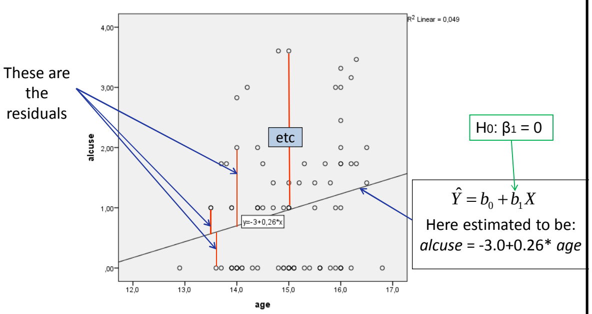
Residual ε: a random variable from a normal distribution with unknown, constant variance σ^2, independent from the value of X

Here: we assume the mean alcohol use values for fixed age values are on a straight line and the individual observations are assumed to be normally distributed around these means (random residual)

Linear regression analysis:
estimate β_0, β_1 by b_0, b_1: find the line which is ``closest'' to the observed data points (ordinary least squares)

Intermezzo
The linear regression model revisited (3)

Example: cross-sectional alcohol-data with best fitted straight line



$$\hat{Y} = b_0 + b_1 X$$

Here estimated to be:
alcuse = -3.0+0.26* *age*

Intermezzo
The linear regression model revisited (4)

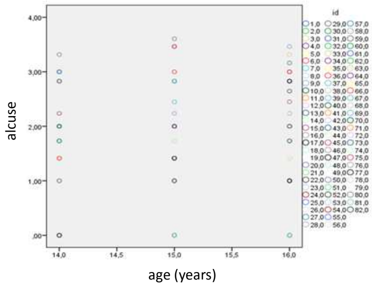
Checking the assumptions made:
- independent observations
- linear relation between Y and X
- normally distributed residuals
  - QQ-plot or PP-plot
- homogeneity of the residual's variance across values of X
  - scatterplot of Zresid against Zpred

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$
$$\varepsilon_i \sim N(0, \sigma^2)$$

Back to our longitudinal data-example...

Investigating change over time
Back to our longitudinal data example

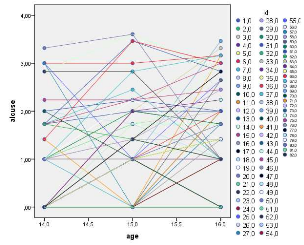
Scatterplot *age-alcuse* for the whole data-set:



Longitudinal data
Plot of whole group

We would like to investigate questions like:

- are there systematic differences between trajectories?
- do these differences increase/decrease?
- does each adolescent follow its own curve?
- what is the effect of COA?



... what about linear regression of alcohol use on age?

Different measurements from one adolescent are related: dependency within observations! Linear regression is no longer an option...

Analysis of longitudinal data
Using summary measures (1)

- To investigate the effect of covariates on the alcohol use of adolescents summary statistics could be investigated
- Choose a summary measure *Y* which reflects a relevant feature of the curve (e.g. the mean, maximum value, time of reaching the maximum, maximal velocity, the last value,...)
- Now there is just one outcome variable (the summary measure) per adolescent: independent observations -> multiple regression analysis!

Advantages:
- simple and easy (can be done using standard techniques)
- provides nice summaries of the data

Disadvantages:
- inefficient use of the whole data
- possible heterogeneity of variance for the summary measure

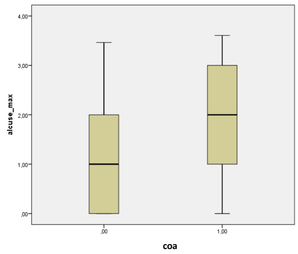
Analysis of longitudinal data
Using summary measures (2)

Example: for each adolescent we take the maximum value of alcohol use *alcuse_max* over the three years:

- Higher median of *alcuse_max* for COA=1 than for COA=0
- Different distributions of two groups: *alcuse_max* much more skewed in COA=0 than in COA=1

(floor-effect due to those who never used alcohol!)

- Does COA affect maximum alcohol use?
  (Mann Whitney/T-test)



We want better use of our data!

16

Analysis of longitudinal data
Summarizing so far...

- Investigating change over time requires multiple (ideally ≥ 3 waves) measurements over time per subject (longitudinal data)
- Linear regression model is not applicable, due to dependency in longitudinal data
- Using summary measures is an option, but it means throwing away information and is limited in answering research questions on change
- Using a cross-sectional data-set instead does not answer research questions on change either

Note: differences between groups of different age ≠ systematic individual change: the highest scoring person at one age need not be the highest scoring person at another age!

17

Analysis of longitudinal data
Introducing the multilevel model for change

We want to expand the linear regression model with several random effects:

mixed effects or multilevel model

random effects & fixed effects — individual level & group level

Enables answers to:

- <u>within-person</u> questions (intra-individual) — Level 1
  How does each person change over time?
  What is each child's rate of development?
- <u>between-person</u> questions (inter-individual) — Level 2
  What predicts differences among people in their change?
  How do these rates vary by child characteristics?

multilevel model for change (linked pair of statistical models)

Analysis of longitudinal data
Introducing the multilevel model for change

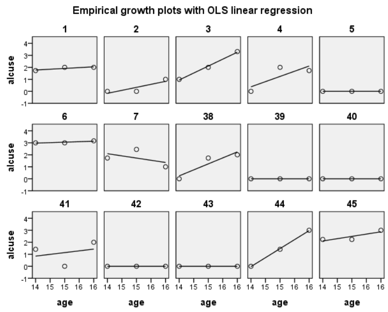
For the remaining lecture-time:

we introduce the multilevel model for change with a simple example, specifying the model and fit it to the data in order to give you a rough idea of what's happening in multilevel modeling

(much more could be told...)

Back to the alcohol-use-data...

Introducing the multilevel model
Exploring individual's growth plots & trajectories



Plotting regression models for each subject to help answer the question:

What population individual growth model might have generated these sample data?

elevation? tilt?

(non-)linear?

NB: ``simpler is better''

Here we choose a linear model

Introducing the multilevel model
The level-1 submodel for individual change

Key assumption: in the population, *alcuse$_j$* is a linear function of child i's *age* on occasion j

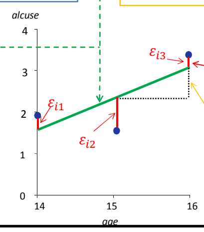
Structural portion: *(hypothesis about) the shape of each person's true trajectory over time*

$$alcuse_{ij} = \pi_{0i} + \pi_{1i} age_{ij} + \varepsilon_{ij}$$

Stochastic portion: *allows for the effects of random error from the measurement of person i on occasion j*

Assumption: $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$

Individual i's hypothesized true trajectory



*π_{0i} is the **intercept** of i's true trajectory (value of alcuse at age=0)*

ε_{i1}, ε_{i2} and ε_{i3} are deviations of i's true trajectory from linearity on each occasion (measurement error)

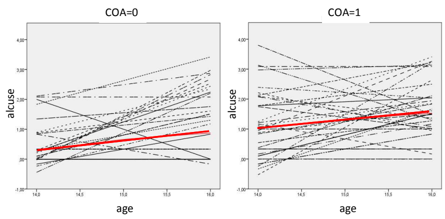
*π_{1i} is the **slope** of i's true change trajectory ("rate of alcuse change")*

i =1, ...,82 (children)
j=1, 2, 3 (measurements)

Introducing the multilevel model
Exploring differences in change across people (inter-individual)

What could be a suitable level-2 model?

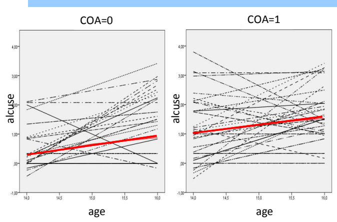
Compare individual trajectories and average change trajectories per group: similarities? differences?



NB: average trajectory need not always have the same shape as individual trajectories!

``curve of averages ≠ average of curves''

Introducing the multilevel model
Exploring differences in change across people (inter-individual)



From these plots:

- children of alcoholic parents (*COA*=1) appear to have higher scores at age 14 (higher intercepts)
- both groups appear to have more or less similar slopes

What is a suitable level-2 model?

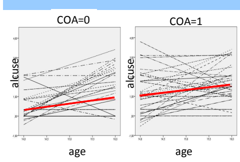
1. Two level-2 submodels : one for each level-1 growth parameter (intercept π_{0i} and slope π_{1i})

2. Each level-2 submodel must specify the relationship between π_{0i} and π_1 and the covariate *COA*

3. Each level-2 submodel should allow individuals with common predictor values (*COA*) to have different individual change trajectories

- We need stochastic variation at level-2, too: each level-2 model will need its own error term,
- ... and we will need to allow for covariance across level-2 errors

Introducing the multilevel model
The level-2 submodels for inter-individual differences in change



Level-2 intercepts
Population average intercept (γ_{00}) and slope (γ_{10}) for COA=0

Level-2 slopes
Effect of COA on intercept (γ_{01}) and on slope (γ_{11})

$$\pi_{0i} = \gamma_{00} + \gamma_{01} COA_i + \zeta_{0i} \quad \text{(intercept)}$$
$$\pi_{1i} = \gamma_{10} + \gamma_{11} COA_i + \zeta_{1i} \quad \text{(slope)}$$

Level-2 residuals
Deviations of each individual's trajectory around the predicted average intercept and slope
(allowing for ``scattering'' of the individual trajectories around the population mean growth trajectories)

The multilevel model for change
Summarizing the total model

	Level:	Predictor(s):	Assumptions:
$alcuse_{ij} = \pi_{0i} + \pi_{1i} age_{ij} + \varepsilon_{ij}$	1	*age*	$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$
$\pi_{0i} = \gamma_{00} + \gamma_{01} COA_i + \zeta_{0i}$ $\pi_{1i} = \gamma_{10} + \gamma_{11} COA_i + \zeta_{1i}$	2	*COA*	$\begin{bmatrix} \zeta_{0i} \\ \zeta_{1i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{bmatrix} \right)$

Here, there are 8 unknown parameters to be estimated:
- 4 fixed effects (level-2 intercepts and slopes) $\gamma_{00}, \gamma_{01}, \gamma_{10}, \gamma_{11}$
- 3 between-person covariances: $\sigma_0^2, \sigma_1^2, \sigma_{01}$ (belonging to the random effects π_{0i} and π_{1i})
- 1 within- person variance: σ_ε^2 (belonging to ε_{ij})

beyond the scope of today's lecture

Introducing the multilevel model
Fitted multilevel model for change: fixed effects (γ_{00}, γ_{01}, γ_{10}, γ_{11})

$$alcuse_{ij} = \pi_{0i} + \pi_{1i} age_{ij} + \varepsilon_{ij}$$
$$\pi_{0i} = \gamma_{00} + \gamma_{01} COA_i + \zeta_{0i}$$
$$\pi_{1i} = \gamma_{10} + \gamma_{11} COA_i + \zeta_{1i}$$

Initial status (``alcuse at age 0'') for the average non-COA adolescent is -3.8

For the average COA-adolescent, it is 1.4 higher (at age 0)

(difference in initial alcuse between COA-groups)

Fitted model for initial status: $\hat{\pi}_{0i} = -3.8 + 1.4 * COA_i$

Fitted model for rate of change: $\hat{\pi}_{1i} = 0.29 - 0.05 * COA_i$

Annual rate of change (slope) for the average non-COA adolescent is 0.29

For the average COA-adolescent, it is 0.05 lower (non significant)

(difference in slope between COA-groups)

Introducing the multilevel model
Constructing prototypical fitted growth trajectories

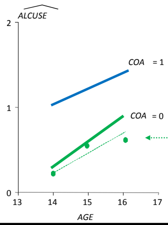
$$\hat{\pi}_{0i} = -3.8 + 1.4 * COA_i$$
$$\hat{\pi}_{1i} = 0.29 - 0.05 * COA_i$$

For *COA*=0 we get: $\hat{\pi}_{0i} = -3.8$, $\hat{\pi}_{1i} = 0.29$

For *COA*=1 we get: $\hat{\pi}_{0i} = -3.8 + 1.4 * 1 = -2.4$, $\hat{\pi}_{1i} = 0.29 - 0.05 * 1 = 0.24$

Substitute these estimated growth parameters into the level-1 model to get fitted growth trajectories:

$$when\ COA_i = 1: \quad \hat{Y}_{ij} = -2.4 + 0.24 * age$$
$$when\ COA_i = 0: \quad \hat{Y}_{ij} = -3.8 + 0.29 * age$$



dotted line: individual estimated trajectory for one child i (randomly deviation from the bold green curve due to ζ_{0i}, ζ_{1i})

green dots: actual observed values of alcuse for child i (randomly scattered around the dotted green line due to ε_{ij})

The multilevel model for change
Combining the levels: rewriting the model

Specification in submodels (level-1 and level-2)

$$\pi_{0i} = \gamma_{00} + \gamma_{01} COA_i + \zeta_{0i} \qquad \pi_{1i} = \gamma_{10} + \gamma_{11} COA_i + \zeta_{1i}$$
$$Y_{ij} = \pi_{0i} + \pi_{1i} age_{ij} + \varepsilon_{ij}$$

... rewriting ...

$$Y_{ij} = (\gamma_{00} + \gamma_{01} COA_i + \zeta_{0i}) + (\gamma_{10} + \gamma_{11} COA_i + \zeta_{1i}) * age_{ij} + \varepsilon_{ij}$$

The composite specification:

$$Y_{ij} = [\gamma_{00} + \gamma_{10} age_{ij} + \gamma_{01} COA_i + \gamma_{11} (COA_i * age_{ij})]$$
$$+ [\zeta_{0i} + \zeta_{1i} age_{ij} + \varepsilon_{ij}]$$

Complex residual: values change with time now and are autocorrelated (this is not regular OLS regression anymore!)

The composite specification shows how *alcuse* depends on:
- the level-1 predictor *age* and the level-2 predictor *COA* as well as
- the cross-level interaction term, *COA*age* , i.e. the effect predictor *age* differs by the levels of predictor *COA*

Some final remarks

- A lot more need to be considered in the context of multilevel models, such as:
  - unbalanced/missing data
  - time-dependent covariates
  - other correlation structures/model designs
  - various estimation methods
  - model building
- Similar modelling techniques exist for different types of outcome variables
- Most major statistical software packages can handle these models
- This abundance of possibilities can also be a pitfall: these models are complex and applying them correctly is a challenge

A selection of books and courses

- Snijders & Bosker: Multilevel Analysis. An introduction to basic and advanced multilevel modeling (London, 1999, 2011)
- Verbeke & Molenberghs: Linear mixed models for longitudinal data (New York, 2000)
- Singer & Willet: Applied longitudinal data analysis. Modeling change and event occurence (Oxford, 2003)
- Pinheiro & Bates: Mixed effects models in S and S-plus (New York, 2000)

Courses offered yearly from our unit:
- Mixed models for clustered data
- Applied longitudinal data analysis

Next Help! Statistics! Lunchtime Lecture

Propensity Scoring

Christine zu Eulenburg

December 12, 2017

Room 16