Analysing longitudinal data with hierarchical linear models and identifying subgroups in prevention research

Workshop at 6. EUSPR, Ljubliana, October 21, 2015

Ferdinand Keller
Department of Child and Adolescent Psychiatry and Psychotherapy, University of Ulm

Time schedule (optional)

9:30 – 10:40	Introductory examples and basic concepts of multi-level modeling of - Cross-sectional data - Longitudinal data Analysis of longitudinal data with multi-level models - Differences to ANOVAR - Some mixed models
10:40 – 11:00	*** Coffee break ***
11:00 – 12:00	- Applications and 'how to do it' in SAS and SPSS - Interpretation of results and critical issues
12:00 – 13:00	*** Lunch ***
13:00 – 14:00	Introduction to growth mixture models (GMM) - Basic concepts of mixture analysis - Mixture analysis with longitudinal data - Applications and 'how to do it' in M*plus*
14:00 – 14:20	*** Coffee break ***
14:20 – 16:00	- Potential pitfalls, caveats, 'reality' of latent classes General discussion, questions, literature and software

Overview

1) Standard approaches to analyse longitudinal data
2) Why multi-level analysis?
3) Hierarchical Linear Model (HLM), multi-level model
4) Latent class growth and growth mixture models
5) Optional: other approaches (survival anaylsis, time series analysis)

Research Questions about Change

- What is the *nature* of change over time, on average? For example, is change linear, curvilinear, nonlinear, discontinuous, etc.?
- How do individuals vary with respect to change over time? For example, do all individuals have linear change but vary in terms of the magnitude of their change coefficients? Or, do individuals differ in terms of the nature of change, e.g., do some individual have linear change while others have curvilinear change?
- What are the effects of risk and protective factors and the intervention on individual differences in the change process?
- How are individual differences in the change process predictive of subsequent or distal outcomes?

from: Masyn & Muthen

Overview: methods for longitudinal data I

1) Standard approaches (well known (more or less))

Regression analysis:

multiple (example: prediction of „outcome“-values)
logistic (example: prediction of „outcome“-groups,
e.g., intervention successful vs. not)

Analysis of Variance with repeated measures (ANOVAR)

compare mean values of group/subgroups over time

MANOVA

compare mean values of group/subgroups over time and
model the covariance/correlation structure over time

Overview: methods for longitudinal data II

3) Hierarchical linear model (HLM), multi-level-model

cross-sectional: e.g. students „nested“ in classes; persons living in neighbourhoods; lenght of stay in hospital (patients on wards, or treated by same therapist);
in general: take into account clustering

longitudinal: growth curves; random coefficients models

4) Latent class growth models, growth mixture models (aim: identify subgroups with different courses)

growth mixture model (GMM): e.g. trajectories of binge drinking in adolescents; course of delinquency; different response to treatment of depression or prevention intervention

latent class growth model (Nagin et al.): special case of GMM (no within class variation)

2) Why multi-level analysis?

cross-sectional:

students „nested“ in classes tend to be more similar to each other than to students in other classes;

analogue:

persons living in the same neighbourhoods;
patients treated in the same hospital (or ward, or therapist);

in general: variances are smaller than without clustering.

thus: take into account clustering for evaluation of effects

Why multi-level: Example 1a
(work satisfaction and responsibility)

Tabelle 19.1 Datenbeispiel für den Zusammenhang zwischen Verantwortung und Arbeitszufriedenheit

Firma	Verantwortung (X)	Arbeitszufriedenheit (Y)
A ($i = 1$)	$x_{11} = 1$	$y_{11} = 8$
	$x_{21} = 2$	$y_{21} = 7$
	$x_{31} = 2$	$y_{31} = 8$
	$x_{41} = 3$	$y_{41} = 9$
	$x_{51} = 5$	$y_{51} = 9$
B ($i = 2$)	$x_{12} = 4$	$y_{12} = 4$
	$x_{22} = 4$	$y_{22} = 6$
	$x_{32} = 5$	$y_{32} = 6$
	$x_{42} = 6$	$y_{42} = 7$
	$x_{52} = 6$	$y_{52} = 8$
C ($i = 3$)	$x_{13} = 5$	$y_{13} = 2$
	$x_{23} = 8$	$y_{23} = 1$
	$x_{33} = 8$	$y_{33} = 2$
	$x_{43} = 8$	$y_{43} = 3$
	$x_{53} = 9$	$y_{53} = 3$

From: Eid, Gollwitzer & Schmitt (2010): Statistik und Forschungsmethoden. Beltz-Verlag

Why multi-level: Example 1b
(work satisfaction and responsibility)

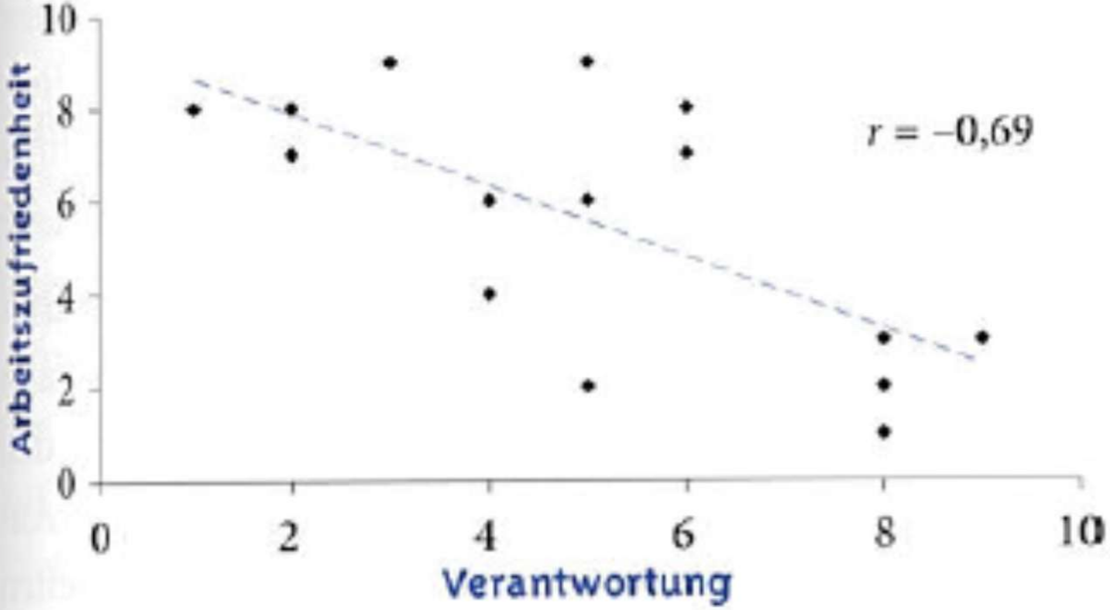


Abbildung 19.1 Streudiagramm für das Datenbeispiel aus Tabelle 19.1: Regression über alle Messwerte hinweg

From: Eid, Gollwitzer & Schmitt (2010): Statistik und Forschungsmethoden. Beltz-Verlag

Why multi-level: Example 1c
(work satisfaction and responsibility)

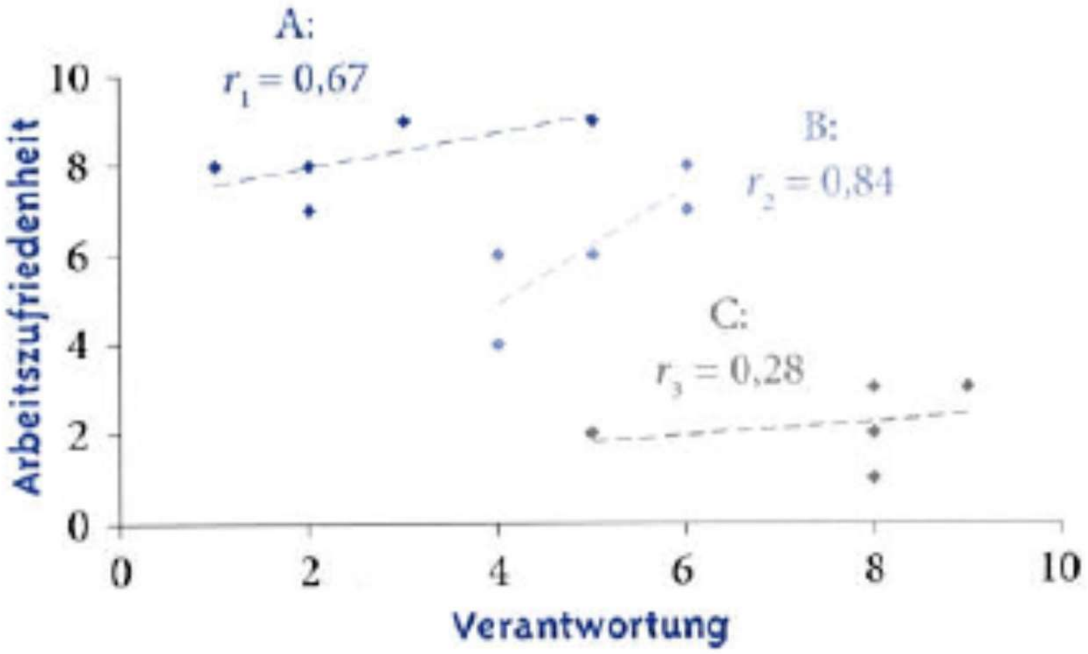


Abbildung 19.2 Streudiagramm für das Datenbeispiel aus Tabelle 19.1: Regression für jede Firma getrennt

From: Eid, Gollwitzer & Schmitt (2010): Statistik und Forschungsmethoden. Beltz-Verlag

Why multi-level: Example 1d
(work satisfaction and responsibility)

a) Decrease of satisfaction with increasing responsibility?

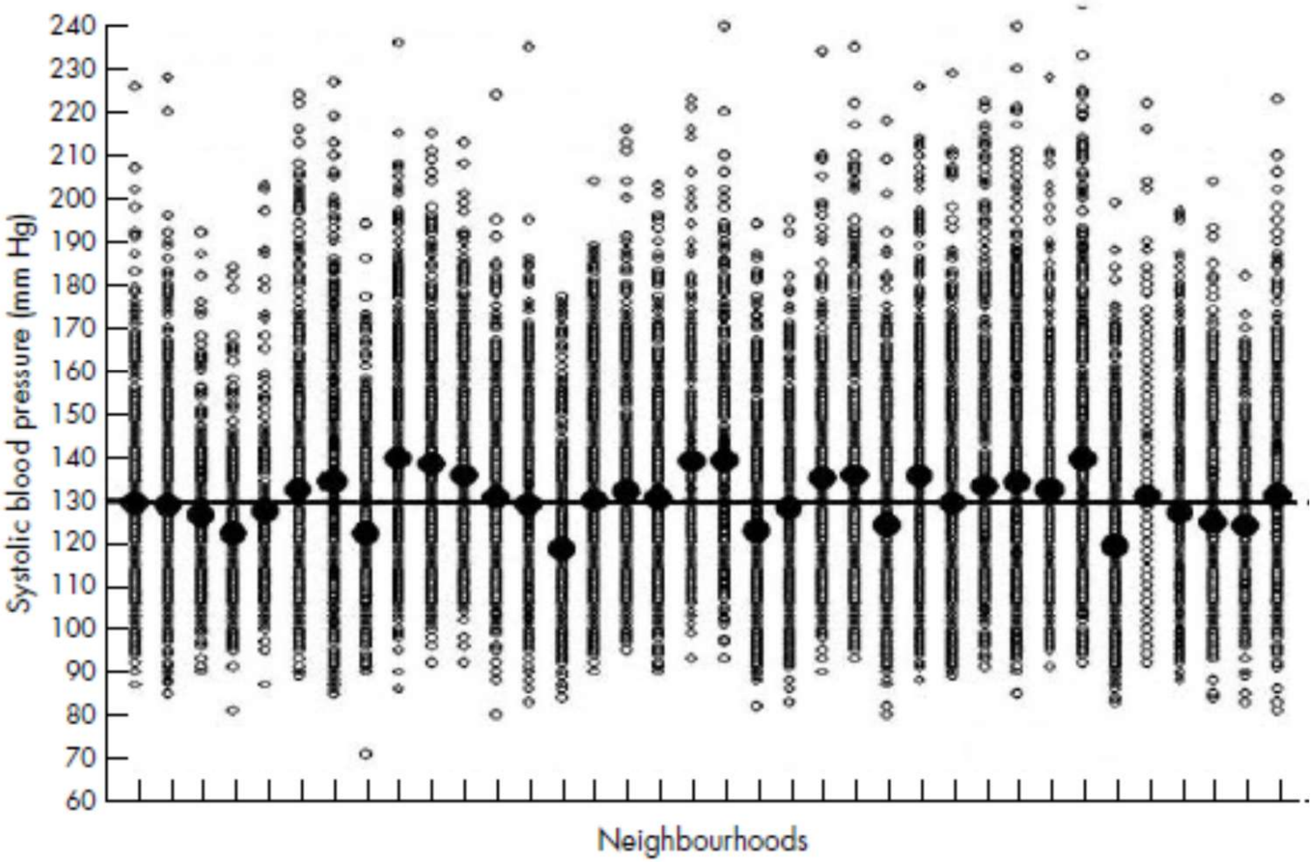
b) Positive correlation within each company?

Ignoring the hierarchical structure (level 2-units) leads to ecological fallacy;

here: Effect on level-2 (company) interpreted on level-1 (individuals)

Problem in this example: differences between companies are big, and explain most of the variance

Why multi-level? Example 2 a
(blood pressure in neighborhoods (simulated data!)



A brief conceptual tutorial of multilevel analysis in social epidemiology: linking the statistical concept of clustering to the idea of contextual phenomenon

Juan Merlo, Basile Chaix, Min Yang, John Lynch, Lennart Råstam

J Epidemiol Community Health 2005;59:443–449. doi: 10.1136/jech.2004.023473

Why multi-level? Example 2 b
(blood pressure in neighborhoods (simulated data!)

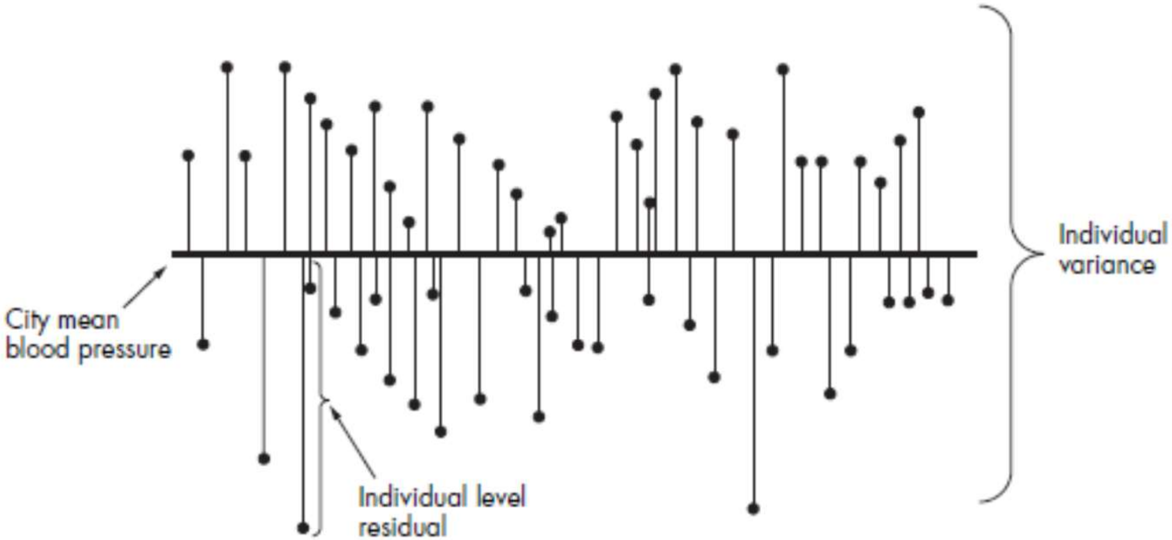


Figure 3 Single level individual information. This figure represents the distribution of individual SBP in the population of the city when we have only single level individual based information. The fact that people are grouped within neighbourhoods is neglected, as we only have individual level data. In this figure the length of the thin vertical line from the black spot to the thick horizontal line represents the individual differences in blood pressure compared with whole city mean (the individual level residuals). The individual variance in single level individual studies is an average summary of these differences. In single level individual analysis we consider all information as if it were at the individual level neglecting possible neighbourhood components.

A brief conceptual tutorial of multilevel analysis in social epidemiology: linking the statistical concept of clustering to the idea of contextual phenomenon

Juan Merlo, Basile Chaix, Min Yang, John Lynch, Lennart Råstam

J Epidemiol Community Health 2005;**59**:443–449. doi: 10.1136/jech.2004.023473

Why multi-level? Example 2 c
(blood pressure in neighborhoods (simulated data!)

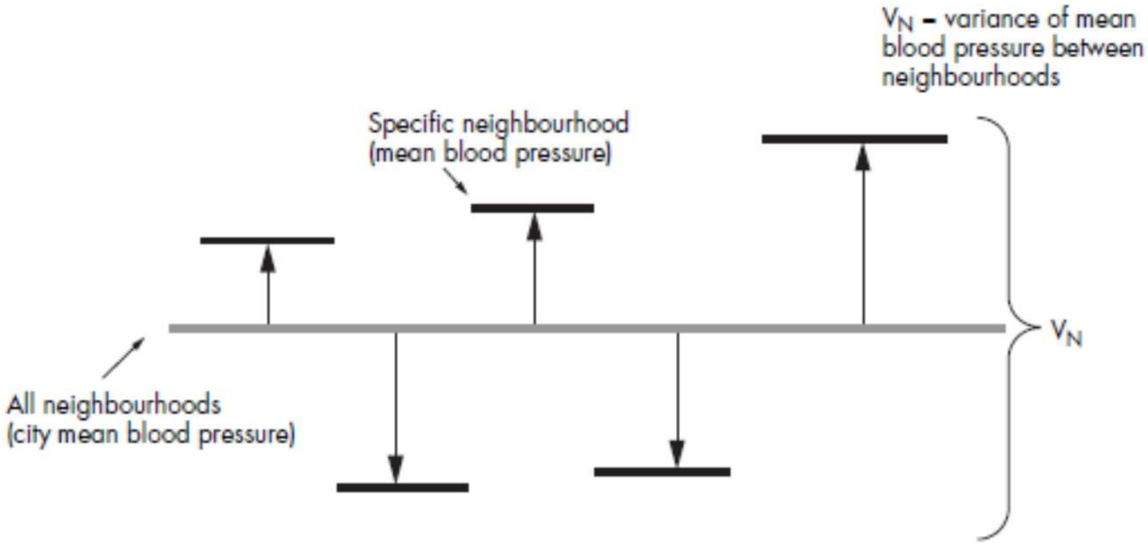


Figure 4 Single level ecological information. In this figure all individual SBP values are aggregated at the neighbourhood level to obtain the neighbourhood mean. We can distinguish differences between the mean blood pressure of each neighbourhood and the mean blood pressure of the whole city (the neighbourhood residuals). These residuals are represented by thick black horizontal lines at the top of a pillow. The neighbourhood variance is a summary of the differences between neighbourhoods. We are unable to observe differences between people (variation in blood pressure within neighbourhoods). In single level ecological analysis we consider all information as if it were at the neighbourhood level neglecting individual components.

A brief conceptual tutorial of multilevel analysis in social epidemiology: linking the statistical concept of clustering to the idea of contextual phenomenon

Juan Merlo, Basile Chaix, Min Yang, John Lynch, Lennart Råstam

J Epidemiol Community Health 2005;59:443–449. doi: 10.1136/jech.2004.023473

Why multi-level? Example 2 d (blood pressure in neighborhoods (simulated data!)

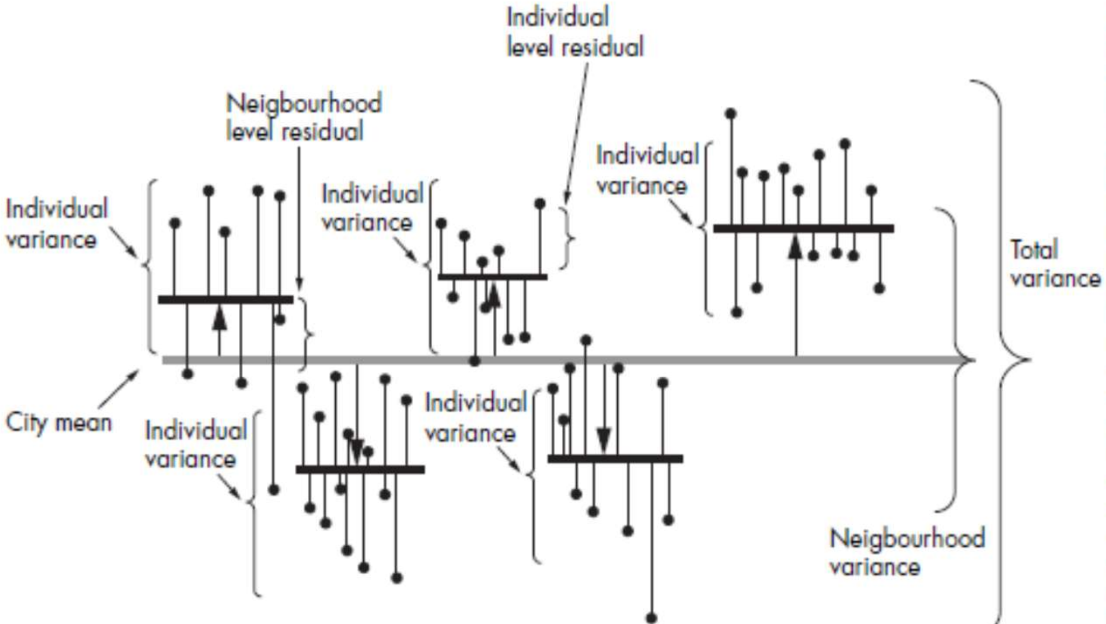


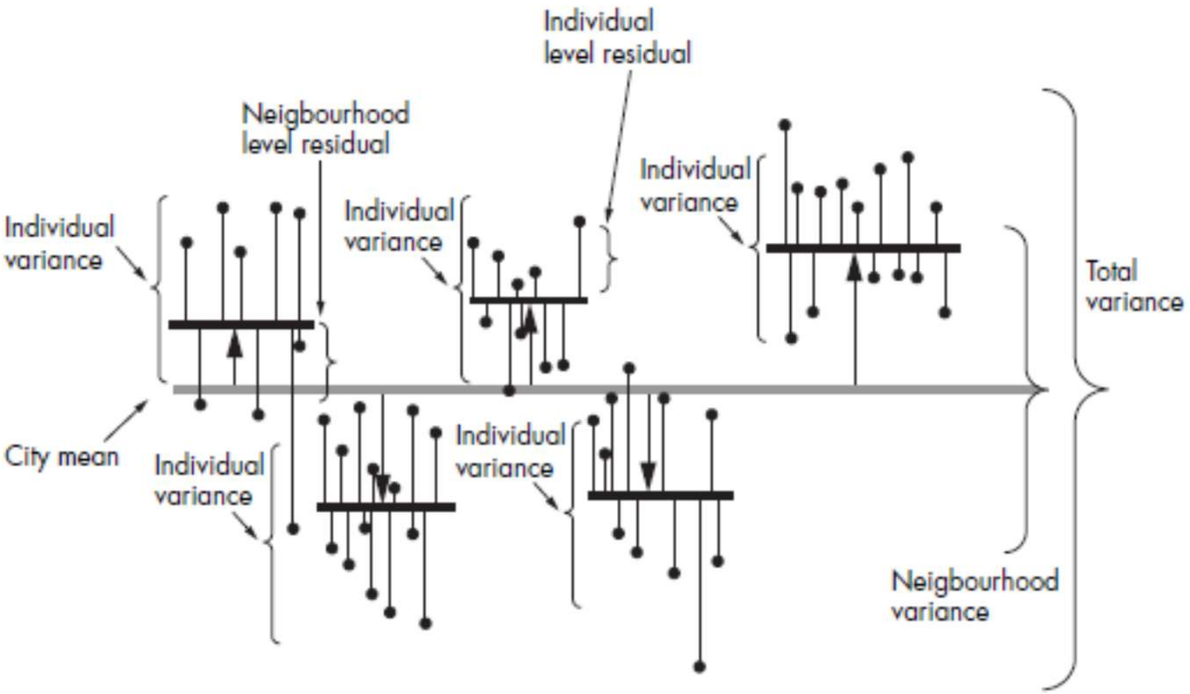
Figure 1 Multilevel information. In this figure the neighbourhood residuals are represented by the length of the pillows between the city SBP mean, represented by a grey colour, and the neighbourhood SBP means represented by thick black horizontal lines. The individual residuals are represented by the length of the vertical lines between the neighbourhood means and the individual SBP values represented by black circles at the top of thin lines. In this figure we do not have any explanatory variable (that is, this figure corresponds to an "empty" model) as we are only interested in analysing how individual blood pressure differences are partitioned in a variability that exists between people from the same neighbourhood and a variability that exists between neighbourhoods. In this figure we can imagine that the neighbourhood means (short thick lines) pull up or pull down all the individual SBP values belonging to the same neighbourhood, even if individual level variability remains within neighbourhoods. The mathematical expression of the intraclass correlation can be visually understood in figure 1. Figure 1 is a graphic combination of figures 3 and 4.

A brief conceptual tutorial of multilevel analysis in social epidemiology: linking the statistical concept of clustering to the idea of contextual phenomenon

Juan Merlo, Basile Chaix, Min Yang, John Lynch, Lennart Råstam

J Epidemiol Community Health 2005;59:443–449. doi: 10.1136/jech.2004.023473

Why multi-level? Example 2 e
(blood pressure in neighborhoods (simulated data!)



How much of the variance is explained by level 2 (neighborhood)?

Intraclass-coefficient (ICC) = Var(Neighb.) / (Var(total) = 0.08

Interpretation: 8% of the variance is explained by level 2

Why multi-level? Conclusions

1) Ignoring cluster effects leads to ‚ecological fallacy' (the more, the higher the ICC)

hence: take account of the hierarchical structure and of interpretation on what level.

Example work satisfaction: explaining level 2 variable, e.g. financial situation of the company, was ignored

also possible: Interaction level 1 and level 2 („cross-level")
example: „big fish little pond"

2) Variances are underestimated (depending on ICC), i.e., standard errors in the denominator of a test statistic are „too small", resulting in „too many" significant effects.

3) Power analysis: less power in clustered designs, i.e., sample size must be increased adequately
(use „variance inflation factor" or via simulations)

Multi-level in longitudinal designs

Transfer from cross-sectional to longitudinal. Now:

level 1: measurement occasions of a person over time

level 2: person

TABLE 6.1

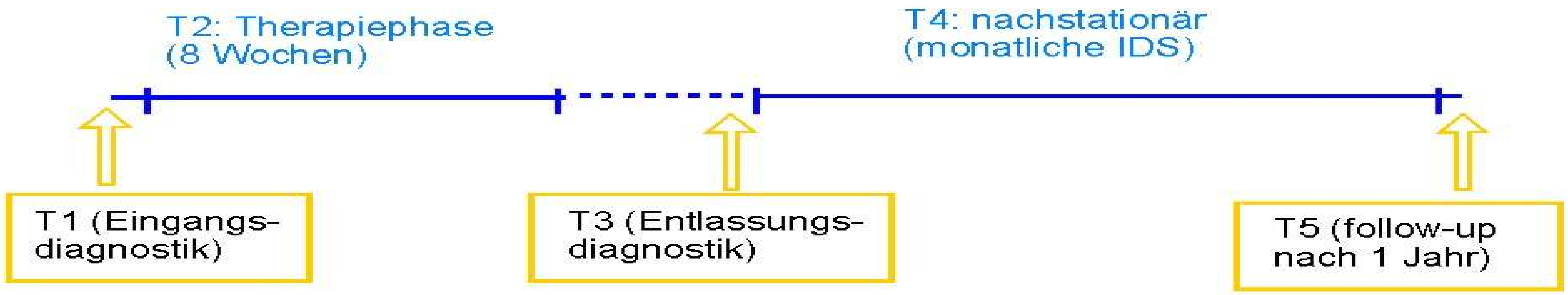
Examples of Longitudinal Data in Different Research Settings

Level of Data		Research Setting: Substance Abuse	Business	Autism Research
Subject (Level 2)	Subject variable (random factor)	College	Company	Child
	Covariates	Geographic region, public/private, rural/urban	Industry, geographic region	Gender, baseline language level
Time (Level 1)	Time variable	Year	Quarter	Age
	Dependent variable	Percent of students who use marijuana during each academic year	Stock value in each quarter	Socialization score at each age
	Time-varying covariates	School ranking, cost of tuition	Quarterly sales, workforce size	Amount of therapy received

From: West/Welch/Galecki (2007). Linear Mixed Models: A practical guide using statistical software. Chapman & Hall

Example: Treatment of depression (Hautzinger / deJong)

(Z. Klin. Psych. 1996)



Study:

Adult depression treatment study (Hautzinger/deJong)

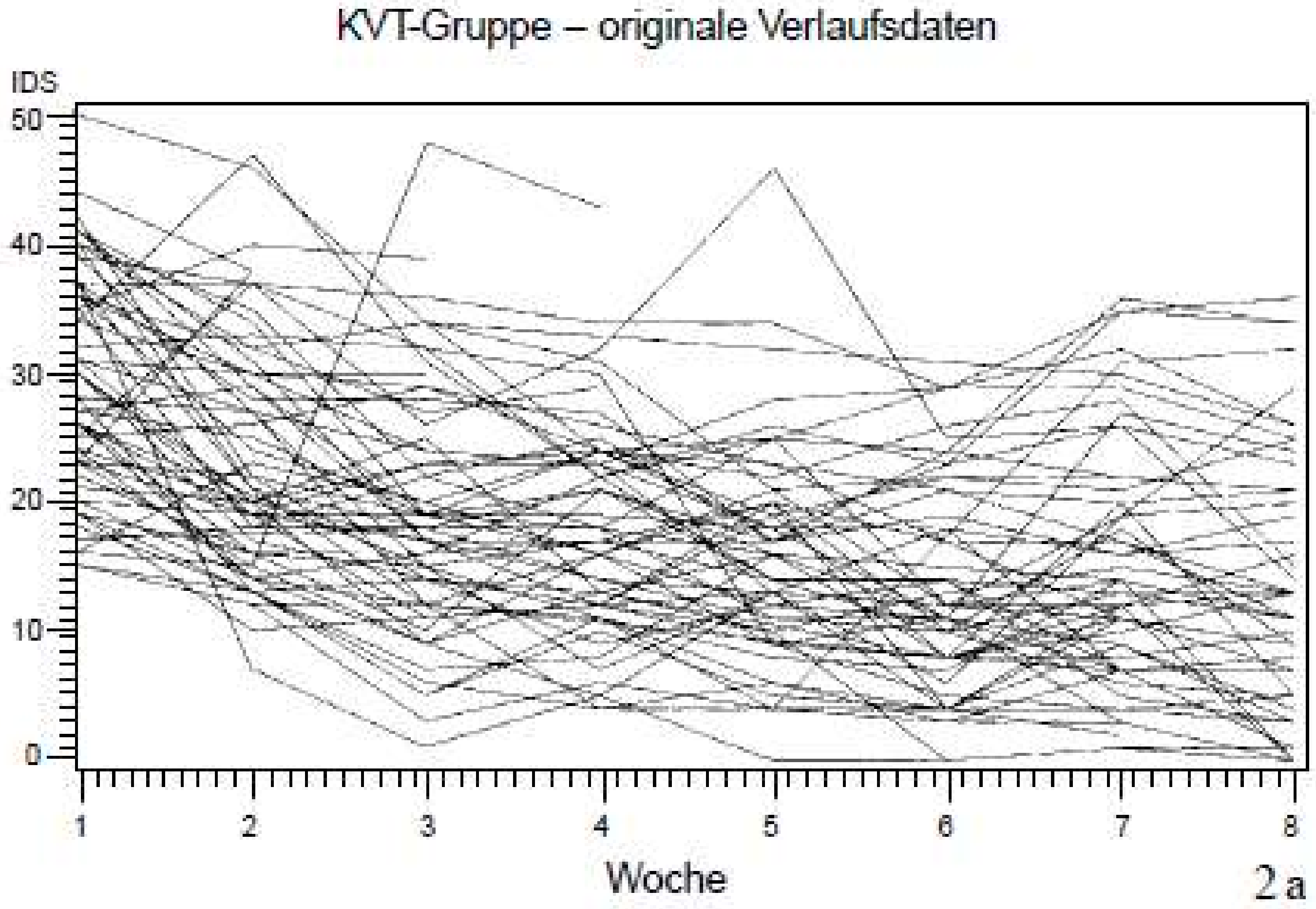
Sample size: n=304

Phase T2 is treatment period: 8 weeks

primary outcome: Inventory of depressive symptoms (IDS), assessed weekly (t=8)

Treatment: medication vs. CBT vs. combination

Therapy study Hautzinger/de Jong
(Raw data, CBT group only)



Aus: Keller, F. (2003): Analyse von Längsschnittdaten: Auswertungsmöglichkeiten mit hierarchischen linearen Modellen. *Zeitschrift für Klinische Psychologie und Psychotherapie, 32,* 51-61

Limitations of ANOVAR

1) „classical“ ANOVAR:

many time points

missing values (--> listwise deletion)

heterogeneity of variances and correlations (correction by Greenhouse/Geisser, Huynh/Feldt)

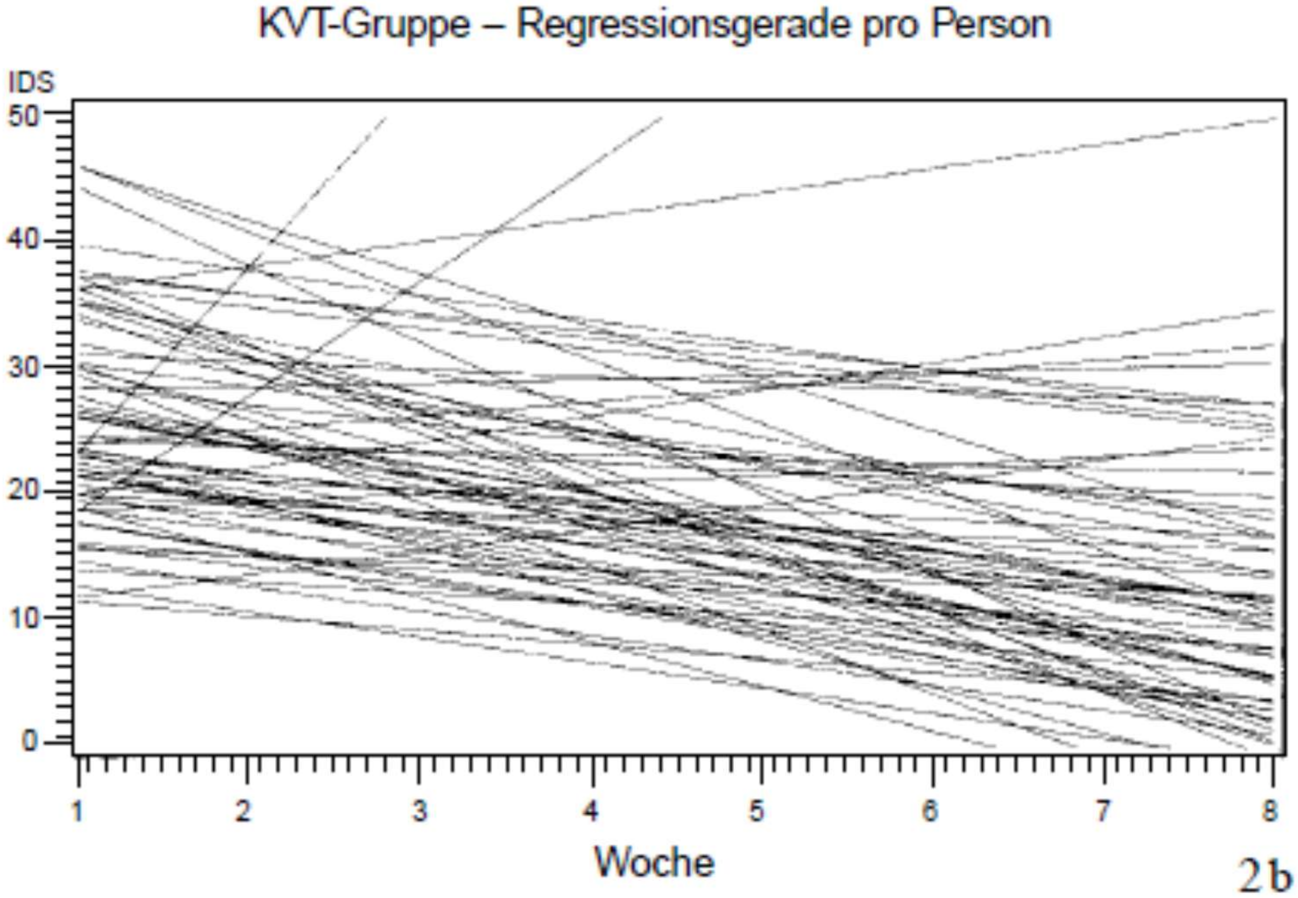
2) In general:

different number of time points per person

unequal distances between time points

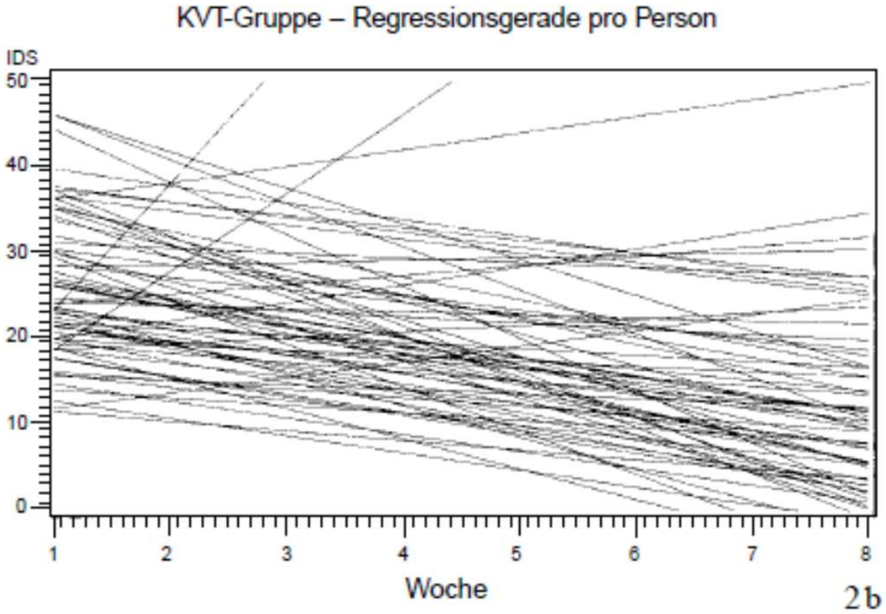
→ Idea: use person-specific regression lines

Therapy study Hautzinger/de Jong
(ordinary regression lines, estimated separately for each person)



Aus: Keller, F. (2003): Analyse von Längsschnittdaten: Auswertungsmöglichkeiten mit hierarchischen linearen Modellen. *Zeitschrift für Klinische Psychologie und Psychotherapie, 32,* 51-61

Therapy study Hautzinger/de Jong
(interpretation and conclusion)



Result: Separate regression lines not helpful, seem „erratic“ in some cases

Expectation for model building and estimation:

- each person belongs to a (sub-)group and their course should be considered and be part of the estimation

- persons with few time points and „unreliable“ course should get less weight

Some practical questions for using HLM

What structure for data file?

Person-level („wide“ format)

Person-period („long“ format)

What software program?

see Singer-Folie (Chap 3, slide 11)

What sample size is needed?

?

How to treat missing data?

See later

Fitting the multilevel model for change to data

Three general types of software options (whose numbers are increasing over time)

Programs expressly designed for multilevel modeling

MLwiN

Multipurpose packages with multilevel modeling modules

Specialty packages originally designed for another purpose that can also fit some multilevel models

aML

Standard structure (person-level data set) („wide“)

The SAS System

Obs	PATNR	miss	IDST2_1	IDST2_2	IDST2_3	IDST2_4	IDST2_5	IDST2_6	IDST2_7	IDST2_8	BDI_Entl	HAM_Entl
1	1101	1	35	26	23	30	.	23	24	21	14	9
2	1102	0	41	34	19	16	12	11	7	1	3	5
3	1103	0	39	19	25	20	20	18	17	18	12	.
4	1104	1	30	34	21	15	9	.	6	10	4	5
5	1105	0	21	20	17	14	11	11	7	7	7	5
6	1106	1	34	42	30	26	24	.	21	16	12	10
7	1107	6	29	21	.	.	.	.	.	.	.	.
8	1108	0	24	29	19	23	14	10	9	5	6	6
9	1110	0	25	37	32	22	26	23	35	34	24	19

Structure required by most statistical packages („long“) (person-period data set)

The SAS System

Obs	PATNR	Woche	IDS_t2	BDI_Entl	HAM_Entl
1	1101	1	35	14	9
2	1101	2	26	14	9
3	1101	3	23	14	9
4	1101	4	30	14	9
5	1101	6	23	14	9
6	1101	7	24	14	9
7	1101	8	21	14	9
8	1102	1	41	3	5
9	1102	2	34	3	5
10	1102	3	19	3	5
11	1102	4	16	3	5

usw.

Programmsyntax (SPSS): long to wide

```
get file 'c:\kidslw.sav'.
list famid birth age.
```

FAMID	BIRTH	AGE
1.00	1.00	9.00
1.00	2.00	6.00
1.00	3.00	3.00
2.00	1.00	8.00
2.00	2.00	6.00
2.00	3.00	2.00
3.00	1.00	6.00
3.00	2.00	4.00
3.00	3.00	2.00

Number of cases read: 9 Number of cases listed: 9

```
casestovars
 /id=famid
 /index = birth
 /drop id kidname wt sex.
list.
```

FAMID	AGE.1.00	AGE.2.00	AGE.3.00
1.00	9.00	6.00	3.00
2.00	8.00	6.00	2.00
3.00	6.00	4.00	2.00

Number of cases read: 3 Number of cases listed: 3


Judith D. Singer - Mozilla Firefox

Datei Bearbeiten Ansicht Chronik Lesezeichen Yahoo! Extras Hilfe

http://gseweb.harvard.edu/~faculty/singer/

Erste Schritte Aktuelle Nachrichten

Search Web Mail Shopping Personals My Yahoo! News Games Travel Finance Answers

HARVARD
GRADUATE SCHOOL OF EDUCATION

Quicklinks:
- Courses
- Downloadable Papers
- Recent Publications
- Older Presentations
- Links

Judith D. Singer

Judith D. Singer (Ph.D., Statistics, Harvard University) is the James Bryant Conant Professor of Education at the Harvard Graduate School of Education and Senior Vice Provost for Faculty Development and Diversity at Harvard University. This website features information about her scholarship, research and teaching. To learn more about her role as Senior Vice Provost, please visit the main Harvard faculty affairs website, www.faculty.harvard.edu.

An internationally renowned statistician and social scientist, Singer's scholarly interests focus on improving the quantitative methods used in social, educational, and behavioral research. She is primarily known for her contributions to the practice of multilevel modeling, survival analysis, and individual growth modeling, and to making these and other statistical methods accessible to empirical researchers.

Singer's wide-ranging interests have led her to publish across a broad array of disciplines, including statistics, education, psychology, medicine, and public health. In addition to writing and co-writing nearly 100 papers and book chapters, she has also co-written three books, including By Design: Planning Better Research in Higher Education and Who Will Teach: Policies that Matter (both published by Harvard University Press). Here is her curriculum vitae.

Her most recent book with longtime collaborator John B. Willett is *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence* (NY: Oxford University Press), for which they received Honorable Mention from the American Publishers Association for the best Mathematics & Statistics book of 2003. Already a classic, ALDA offers an accessible in-depth presentation of two popular statistical methods for analyzing longitudinal data: multilevel modeling of individual change and

Applied Longitudinal Data Analysis

Fertig

A conceptual overview of the multilevel model for change

Key idea: You're building linked statistical models at each of two levels of a hierarchy

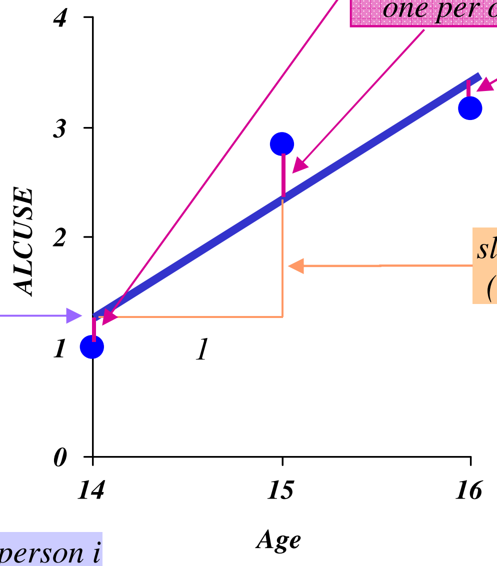
At level-1 (within person)

Model the ***individual change trajectory****, which describes how each person's status depends on time*

Example: ***Changes in alcohol use among teens***

(data for 1 COA from Curran's study of 82 teens interviewed at 14, 15, and 16)

residuals for person i, one per occasion j



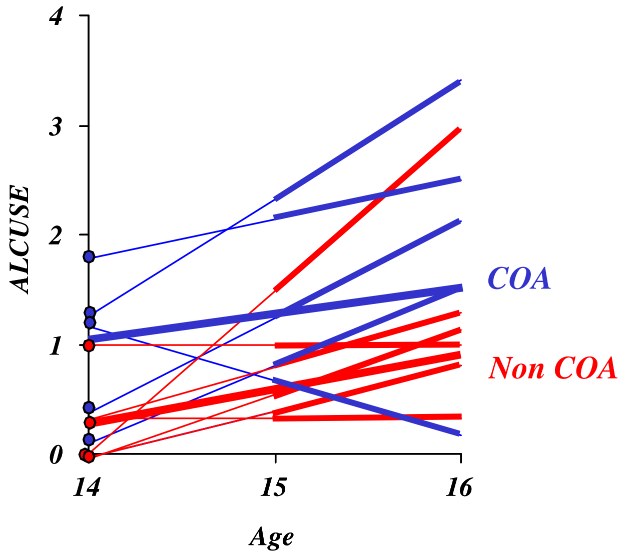
slope for person i ("growth rate")

intercept for person i ("initial status")

$$Y_{ij} = \pi_{0i} + \pi_{1i}(AGE-14)_{ij} + \varepsilon_{ij}$$

At level-2 (between persons)

Model ***inter-individual differences in change****, which describe how features of the change trajectories vary across people*



$$\pi_{0i} = \gamma_{00} + \gamma_{01}COA_i + \zeta_{0i}$$

For intercepts

$$\pi_{1i} = \gamma_{10} + \gamma_{11}COA_i + \zeta_{1i}$$

For slopes

Programmsyntax (SAS Proc MIXED)

Syntax (SAS):

```
PROC MIXED DATA=ids;
   MODEL idst2 = week / solution; * course of all persons;
   RANDOM intercept week /SUB=patnr TYPE=UN G GCorr;
RUN;

PROC MIXED DATA=ids covtest;
   CLASS treat;
   MODEL idst2 = treat week treat*week / solution;
        * time effect + fixed effect treatment;
   RANDOM intercept week /SUB=patnr TYPE=UN G GCorr;
RUN;
```

Results of fitting Model B (the unconditional growth model) to data

$$Y_{ij} = \pi_{0i} + \pi_{1i}(AGE - 14)_{ij} + \varepsilon_{ij} \text{ , where } \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$$

$$\begin{aligned} \pi_{0i} &= \gamma_{00} + \zeta_{0i} \\ \pi_{1i} &= \gamma_{10} + \zeta_{1i} \end{aligned} \quad where \begin{bmatrix} \zeta_{0i} \\ \zeta_{1i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{bmatrix} \right)$$

$$Y_{ij} = \gamma_{00} + \gamma_{10}(AGE - 14)_{ij} + [\zeta_{0i} + \zeta_{1i}(AGE - 14)_{ij} + \varepsilon_{ij}]$$

```
proc mixed data=one method=ml covtest;
   class id;
   model alcuse = age_14/solution;
   random intercept age_14/type=un subject=id;
```

Parameter #1 (intercept) Parameter #2 (age_14)

		Parameter	Model B
Fixed Effects			
Initial status, π_{0i}	Intercept	γ_{00}	0.651*** (0.105)
Rate of change, π_{2i}	Intercept	γ_{10}	0.271*** (0.062)
Variance Components			
Level 1	Within-person	σ_ε^2	0.337*** (0.053)
Level 2	In initial status	σ_0^2	0.624*** (0.148)
	In rate of change	σ_1^2	0.151** (0.056)
	Covariance	σ_{01}	–0.068 (0.070)
	Deviance		636.6
	AIC		648.6
	BIC		663.0

~$p < .10$; *$p < .05$; **$p < .01$; ***$p < .001$

Model B: Unconditional growth model
The Mixed Procedure

```
              Covariance Parameter Estimates

                                Standard      Z
Cov Parm   Subject   Estimate     Error     Value     Pr Z

UN(1,1)    ID          0.6244    0.1481     4.22    <.0001
UN(2,1)    ID        -0.06844    0.07008   -0.98    0.3288
UN(2,2)    ID          0.1512    0.05647    2.68    0.0037
Residual               0.3373    0.05268    6.40    <.0001

                       Fit Statistics

-2 Log Likelihood                 636.6
AIC (smaller is better)           648.6
AICC (smaller is better)          649.0
BIC (smaller is better)           663.1

               Solution for Fixed Effects

                        Standard
Effect       Estimate     Error     DF   t Value   Pr > |t|

Intercept     0.6513    0.1051      81     6.20     <.0001
AGE_14        0.2707    0.06245     81     4.33     <.0001
```

Results of fitting Model C (the uncontrolled effects of COA) to data

$$Y_{ij} = \pi_{0i} + \pi_{1i}(AGE - 14)_{ij} + \varepsilon_{ij} \text{, where } \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$$

$$\begin{aligned} \pi_{0i} &= \gamma_{00} + \gamma_{01} COA_i + \zeta_{0i} \\ \pi_{1i} &= \gamma_{10} + \gamma_{11} COA_i + \zeta_{1i} \end{aligned} \quad where \begin{bmatrix} \zeta_{0i} \\ \zeta_{1i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{bmatrix} \right)$$

$$Y_{ij} = \gamma_{00} + \gamma_{01} COA_i + \gamma_{10}(AGE - 14)_{ij} + \gamma_{11} COA_i * (AGE - 14)_{ij} + [\zeta_{0i} + \zeta_{1i}(AGE - 14)_{ij} + \varepsilon_{ij}]$$

```
proc mixed data=one method=ml covtest;
   class id;
   model alcuse = coa age_14 coa*age_14/solution;
   random intercept age_14/type=un subject=id;
```

		Parameter	Model C
Fixed Effects			
Initial status, π_{0i}	Intercept	γ_{00}	0.316*** (0.131)
	COA	γ_{01}	0.743*** (0.195)
Rate of change, π_{2i}	Intercept	γ_{10}	0.293*** (0.084)
	COA	γ_{11}	−0.049 (0.125)
Variance Components			
Level 1	Within-person	σ_ε^2	0.337*** (0.053)
Level 2	In initial status	σ_0^2	0.488** (0.128)
	In rate of change	σ_1^2	0.151* (0.056)
	Covariance	σ_{01}	−0.059 (0.066)
	Deviance		621.2
	AIC		637.2
	BIC		656.5

~$p < .10$; * $p < .05$; ** $p < .01$; *** $p < .001$

Model C: Uncontrolled effects of COA
The Mixed Procedure

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
UN(1,1)	ID	0.4876	0.1278	3.81	<.0001
UN(2,1)	ID	-0.05934	0.06573	-0.90	0.3666
UN(2,2)	ID	0.1506	0.05639	2.67	0.0038
Residual		0.3373	0.05268	6.40	<.0001

Fit Statistics

-2 Log Likelihood	621.2
AIC (smaller is better)	637.2
AICC (smaller is better)	637.8
BIC (smaller is better)	656.5

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	0.3160	0.1307	80	2.42	0.0179
COA	0.7432	0.1946	82	3.82	0.0003
AGE_14	0.2930	0.08423	80	3.48	0.0008
COA*AGE_14	-0.04943	0.1254	82	-0.39	0.6944

Syntax for SPSS (Procedure MIXED)

```
Mixed
    y with time ccovar
    /print=solution corb
    /fixed = time ccovar time*ccovar
    /random intercept time | subject(id) covtype(un).
```

There is also a good introduction by SPSS:


Technical report

SPSS

Linear Mixed-Effects Modeling in SPSS: An Introduction to the MIXED Procedure

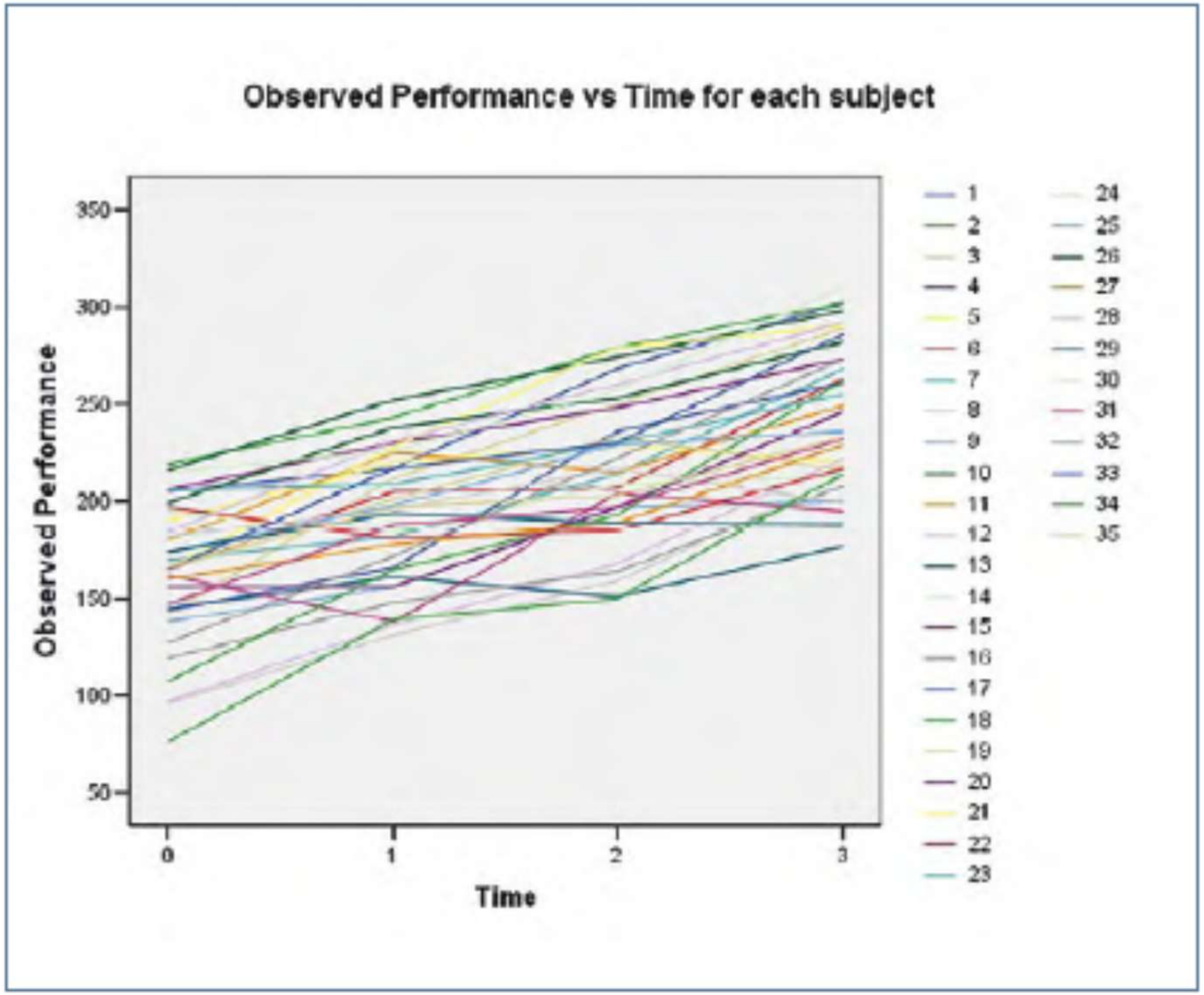


Figure 45

Random coefficient models

In many situations, it is impossible to use a single regression line to describe the behavior of every individual. To account for possible variations between individuals, we can treat the regression coefficients as random variables. This type of model is therefore called the random coefficient model. We typically assume that the regression coefficients have normal distributions.

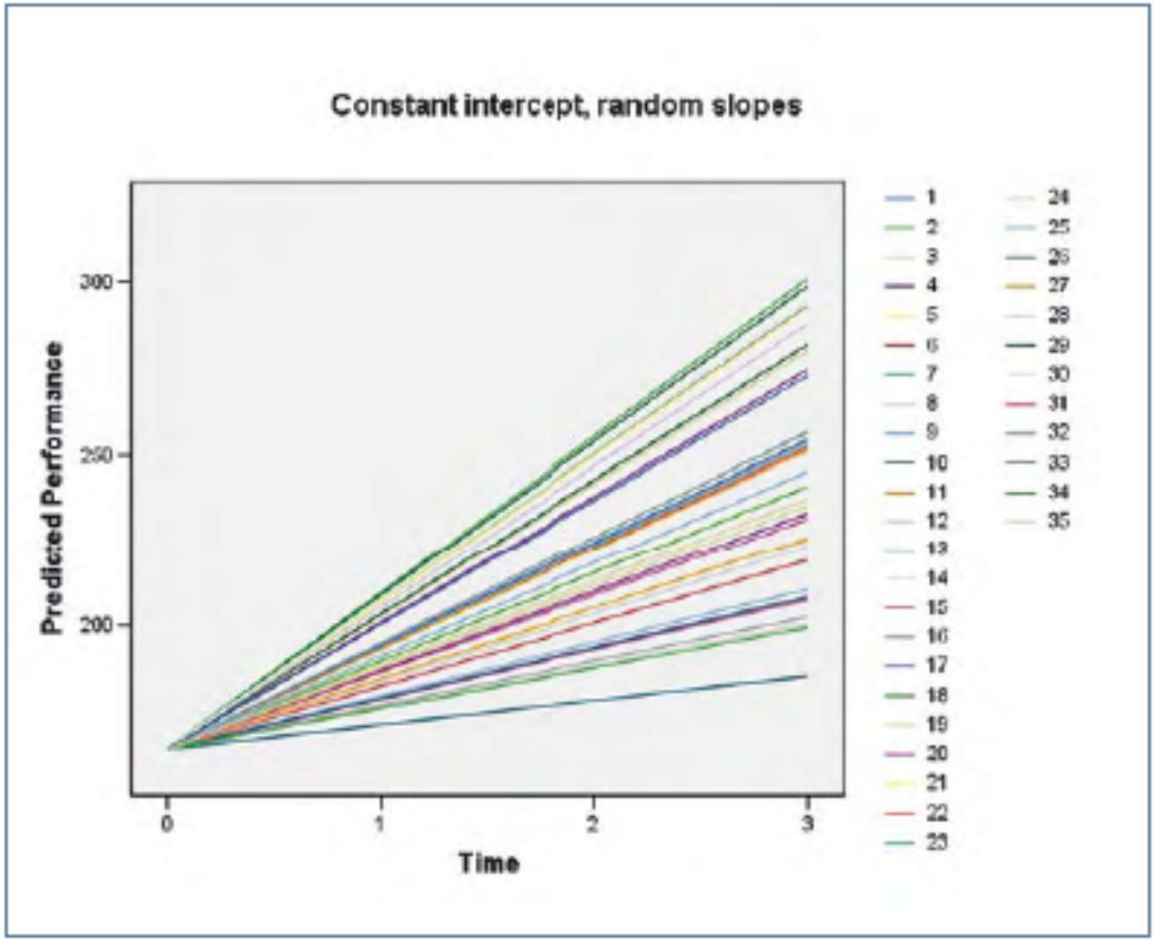
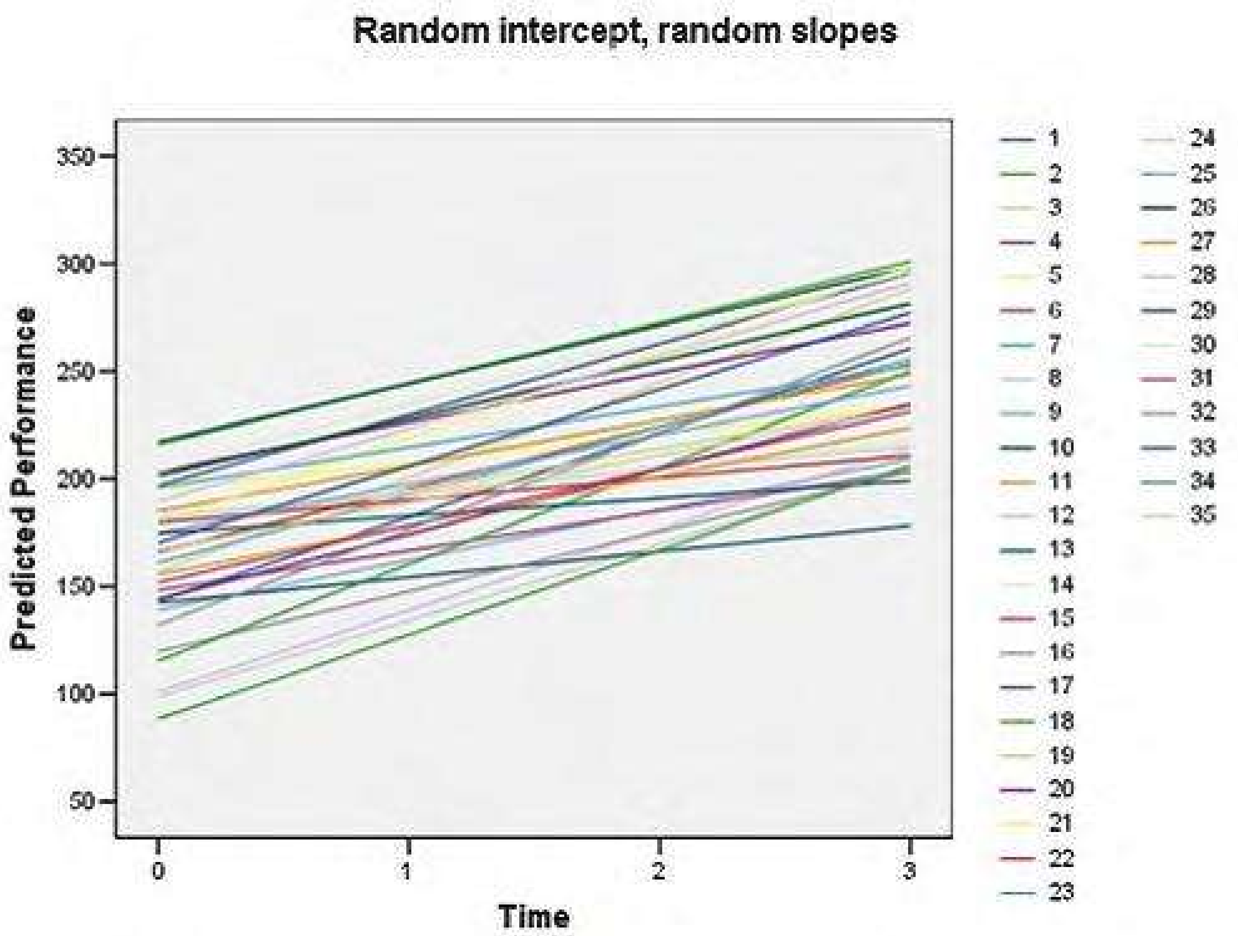

Constant intercept, random slopes
Predicted Performance
200
250
300
Time
0
1
2
3
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35

Figure 51



Model AIC

Random intercept 1304.340

Random slope 1361.464

Random intercept and slope 1274.823

Other models and systematic overview

a) Null model: variation in intercept? (= Cluster effect?)

b) Time effect? (different means over time)

c) Effect of covariate(s)?, e.g, treatment, intervention

d) Variation in slopes? (add time as random effect)

e) Interaction effects?, e.g., time x treatment

- Demonstrated with examples from Hox, J. (200x): Multilevel analysis techniques and applications, chap. 5

Syntax and resulting output via:
http://www.ats.ucla.edu/stat/sas/examples/mlm_ma_hox/chap5.htm

Exkurs: Influence of coding of time

Summary (so far)

Multi-level or hierarchical linear model (HLM)

a) cross-sectional: examples were:

individuals in companies; in neighborhoods;
students in classes; children in families; …..

b) longitudinal: random coefficients models

estimate person specific regression lines (linear, quadratic)
while taking into account group intercepts and slopes

Person specific coefficients (intercept, slope) not of much interest (usually), but their variances

Interpretation (mostly) of fixed effects (treatment, time course, etc.)

Outlook: Models with three levels

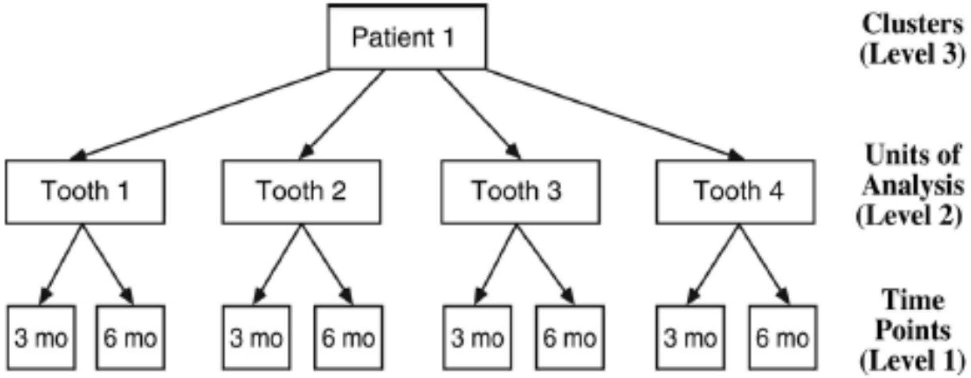
Technically not very difficult. But may be difficult to estimate and interpret.

Multi-level in longitudinal designs: three levels

TABLE 7.1

Examples of Clustered Longitudinal Data in Different Research Settings

Level of Data		Research Setting: Environment	Education	Dentistry
Cluster of Units (Level 3)	**Cluster ID variable (random factor)**	**Plot**	**Classroom**	**Patient**
	Covariates	Soil minerals, tree crown density in the plot	Teacher years of experience, classroom size	Gender, age
Unit of Analysis (Level 2)	**Unit of Analysis ID variable (random factor)**	**Tree**	**Student**	**Tooth**
	Covariates	Tree size	Gender, age, baseline score	Treatment, tooth type
Time (Level 1)	**Time variable**	**Week**	**Marking period**	**Month**
	Dependent variable	Oxygen yield	Test score	Gingival crevicular fluid (GCF)
	Time-varying covariates	Sunlight exposure, precipitation	Attendance	Frequency of tooth brushing



From: West/Welch/Galecki (2007). Linear Mixed Models: A practical guide using statistical software. Chapman & Hall

Research Questions about Change

- What is the *nature* of change over time, on average? For example, is change linear, curvilinear, nonlinear, discontinuous, etc.?
- How do individuals vary with respect to change over time? For example, do all individuals have linear change but vary in terms of the magnitude of their change coefficients? Or, do individuals differ in terms of the nature of change, e.g., do some individual have linear change while others have curvilinear change?
- What are the effects of risk and protective factors and the intervention on individual differences in the change process?
- How are individual differences in the change process predictive of subsequent or distal outcomes?
- How does the change process influence (mediate/moderate) the intervention effect on the distal outcomes?

from: Masyn & Muthen

Overview: methods for longitudinal data II

3) Hierarchical linear model (HLM), multi-level-model

cross-sectional: e.g. classes „nested“ in schools; persons living in neighbourhoods; lenght of stay in hospital (patients on wards); in general: take into account clustering

longitudinal: growth curves; random coefficients models

4) Latent class growth models, growth mixture models (aim: identify subgroups with different courses)

growth mixture model (GMM): e.g. trajectories of binge drinking in adolescents; course of delinquency; different response to treatment of depression

latent class growth model (Nagin et al.): special case of GMM (no within class variation, SAS Proc TRAJ)

„Traditions“ of GMM

From antisocial behaviour/delinquency research:

Latent class growth model **(LCGA)** (Nagin, Tremblay, Jones):

Individual growth trajectories within a class are homogeneous

From latent variable modelling:

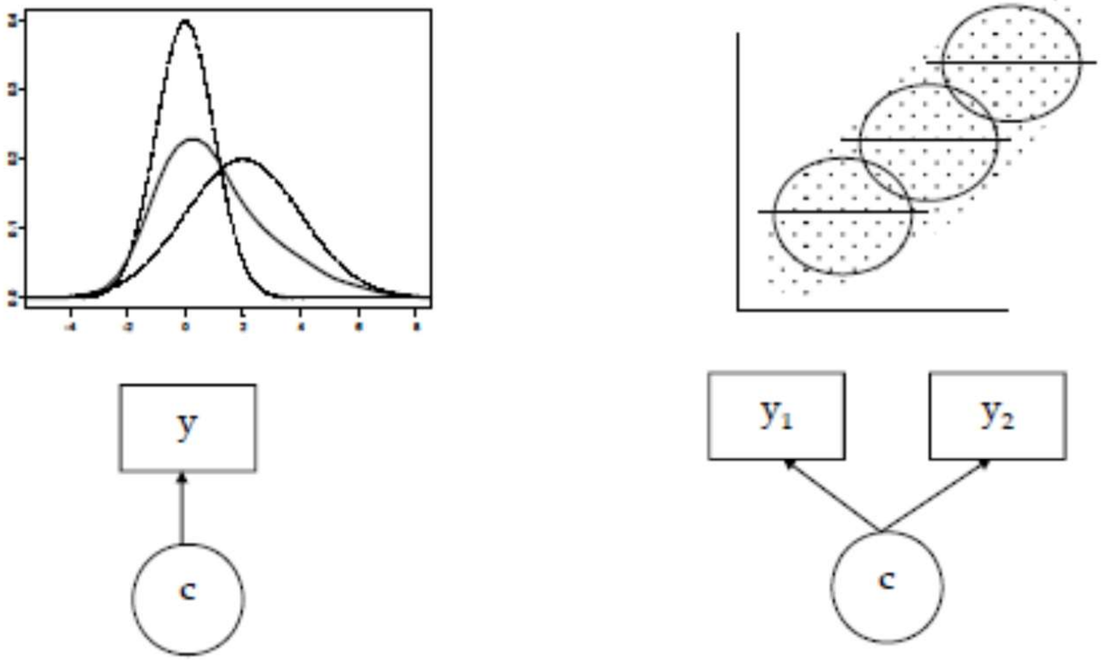
Growth mixture modelling (**GMM**) (Muthen):

Allows for within-class variation (and freeing/fixing of other parameters)

- Heterogeneity in growth is captured through trajectory classes (categorical latent variable) **and** random effects *within* class.

General idea of mixture modeling

MIXTURE MODELING

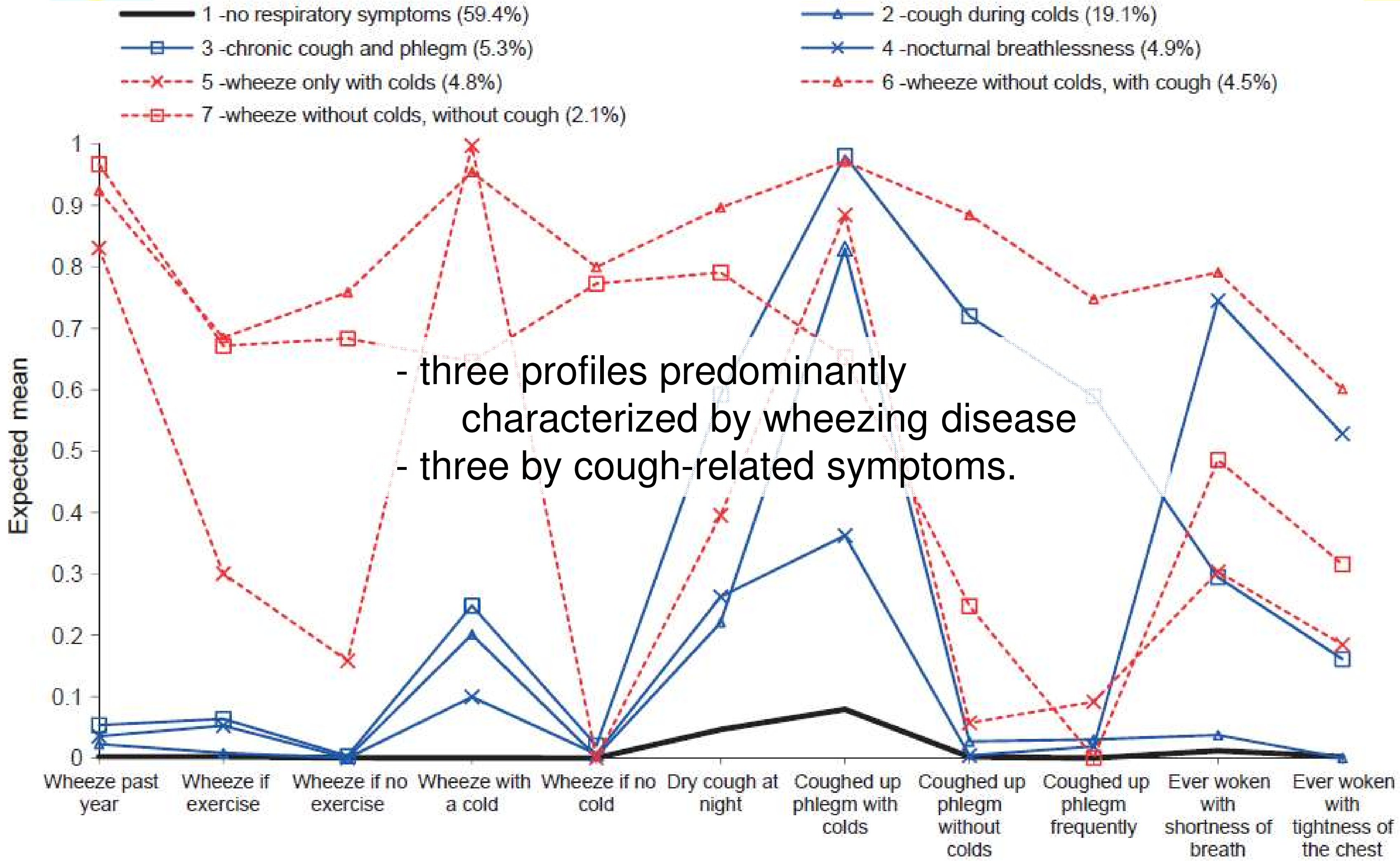


REITOX Academy, EMCDDA

© Masyn & Muthén (13-Mar-2009)

48

Example: Latent class analysis (cross-sect.)



- three profiles predominantly characterized by wheezing disease
- three by cough-related symptoms.

Fig. 1. Estimated prevalences (expected mean) of the respiratory symptoms in, and estimated class sizes (in brackets) of the seven phenotypes identified by latent class analysis.

Weinmayr/Keller/Kleiner et al. (2013). Asthma phenotypes identified by latent class analysis in the ISAAC phase II Spain Study. *Clinical & Experimental Allergy, 43,* 223-232.

Single population vs. unobserved subpopulations

Growth curves as random-effects models:

Growth curve parameters, e.g. intercept and slope, vary across individuals.

However,

- a **single population** with common parameters is assumed, or
- **subgroups are known**, e.g. treatment condition, gender…

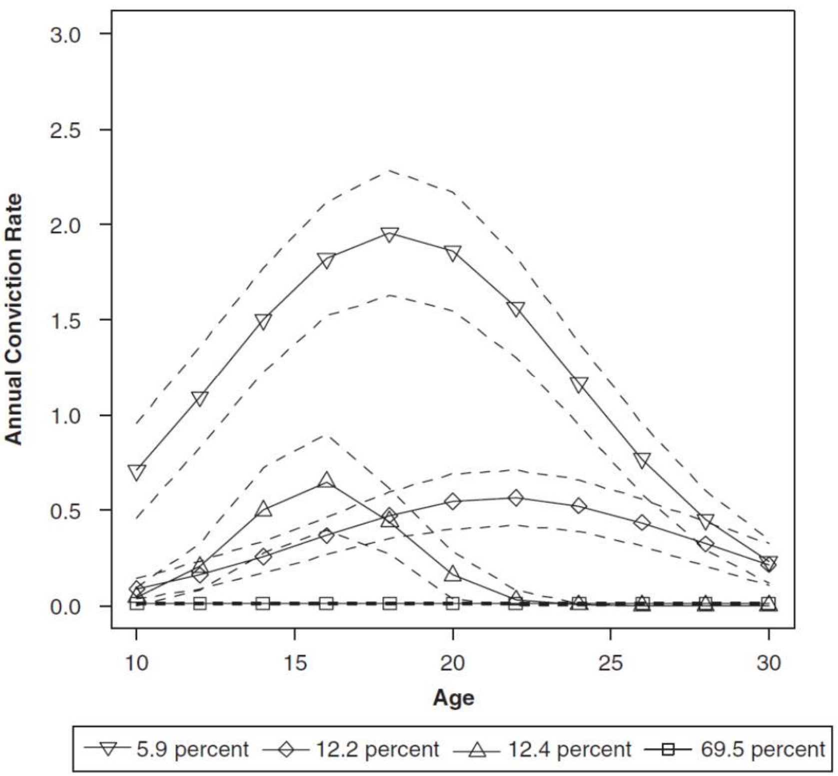
Growth mixture modelling:

allows for differences in growth parameters across **unobserved subpopulations**, resulting in separate growth models for each subpopulation (latent class).

Example: latent class growth model for delinquency

Jones, Nagin / Group-Based Trajectory Modeling 547

Figure 1
Annual Conviction Rate Versus Age: Four-Group Poisson Model



Advances in Group-Based Trajectory Modeling and an SAS Procedure for Estimating Them
Bobby L. Jones and Daniel S. Nagin
Sociological Methods Research 2007; 35; 542

SAS Proc Traj - Home - Mozilla Firefox

http://www.contrib.andrew.cmu.edu/~bjones/index.htm

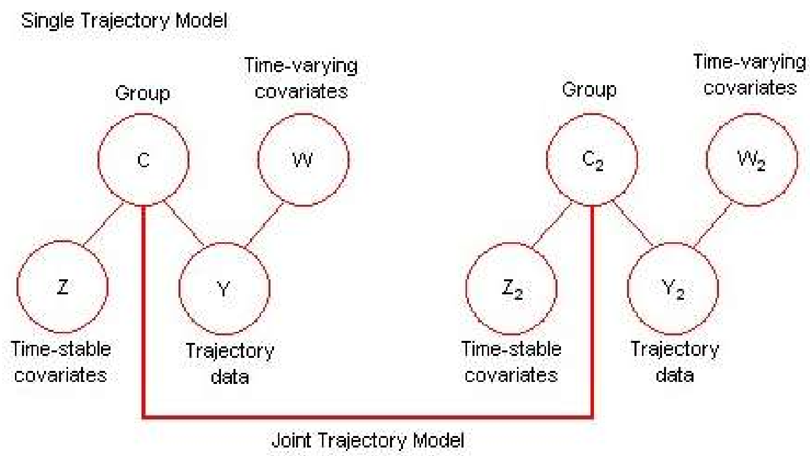
PROC TRAJ

Home | Download | Examples | Documentation

PROC TRAJ is a SAS procedure that fits a discrete mixture model to longitudinal data. The model performs data sequence grouping, with different parameter values for the groups' data distribution. Groupings may identify distinct subpopulations. Alternatively, groups may represent distribution components approximating an unknown (possibly complex) data distribution.

Supported distributions are: censored (or regular) normal, zero inflated (or regular) Poisson, and Bernoulli distributions (logistic model). The censored normal model is useful for psychometric scale data, the zero inflated Poisson model useful for count data with extra zeros, and the Bernoulli model useful for 0/1 data. The model is appropriate for data with average values changing smoothly as a function of the dependent variable (time, age, ...). Some sharp changes can be handled through the inclusion of time dependent covariates.

MODEL STRUCTURE: Data sequences, Y, with similar shapes are grouped in a model-based manner. The probability of group membership can be a function of time stable covariates (risk factors), Z. Time dependent covariates, W, can further influence trajectories with effects differing by group, C. A trajectory model for two sets of dependent variables (joint trajectory modeling) is also supported. The model is illustrated in the figure below.

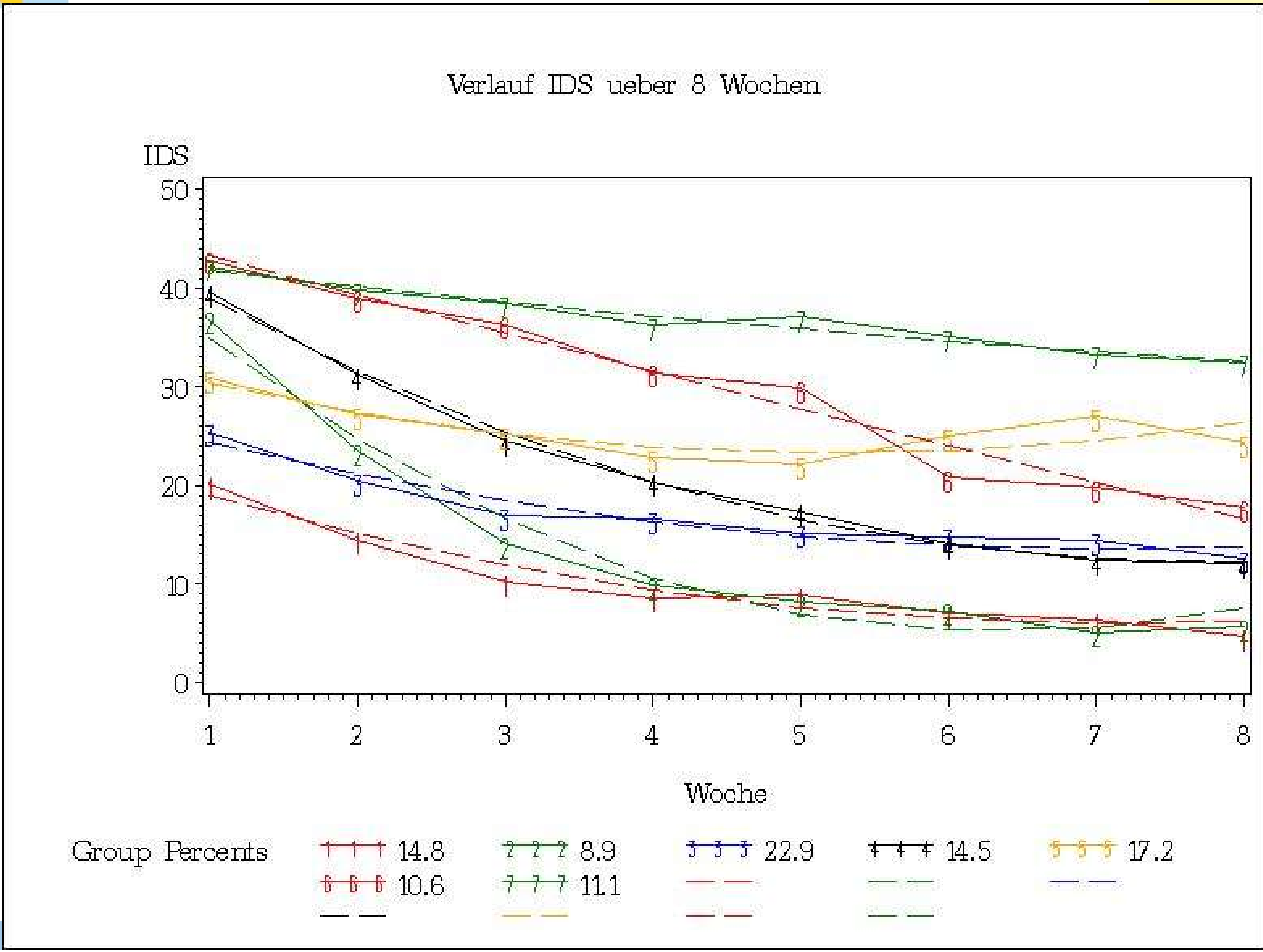


Downloads: Jones, B., Nagin, D., & Roeder, K. "A SAS Procedure Based on Mixture Models for Estimating Developmental Trajectories," Sociol Method Res, 2001, 29: 374-393.

Jones, B. & Nagin, D. "Advances in Group-Based Trajectory Modeling and a SAS Procedure for Estimating Them," submitted

Fertig

Trajectories for 7 classes (LCGA, estimated and sample means)



Note: Current version is 7.3, but Flash for presenting the version no longer allowed…

Latent Class Growth model: Example syntax in M*plus*

```
TITLE: course of IDS over 8 weeks, 7 classes, no random effects;
DATA: FILE IS ids.dat;

VARIABLE: NAMES ARE STUDIE THGRUPPE PATNR BDIT1SUM BDIT3SUM
            HA21SUT3 y1 y2 y3 y4 y5 y6 y7 y8 ;
            USEVAR = y1-y8;
            IDVARIABLE=patnr;
            MISSING ARE ALL (999);
            CLASSES=C(7);

ANALYSIS: TYPE=Mixture missing;
             estimator=ML;
MODEL:
  %OVERALL%
        i s q | y1@0 y2@1 y3@2 y4@3 y5@4 y6@5 y7@6 y8@7;
        i@0; s@0; q@0;     ! Set variances to zero, yields LCGA model ;

OUTPUT: RESIDUAL stand  tech8;
PLOT: Type is plot1 plot2 plot3;  series = y1-y8 (s);
```

Growth mixture model: Programmsyntax in M*plus*

```
TITLE: course of IDS, linear and quadratic, 3 classes, with random effects;
DATA: FILE IS ids.dat;

VARIABLE: NAMES ARE STUDIE THGRUPPE PATNR BDIT1SUM BDIT3SUM
          HA21SUT3 y1 y2 y3 y4 y5 y6 y7 y8 ;
          USEVAR = y1-y8;
          IDVARIABLE=patnr;
          MISSING ARE ALL (999);
          CLASSES=C(3);

ANALYSIS: TYPE=Mixture missing;
          estimator=ML;
MODEL:
 %OVERALL%
     i s q | y1@0 y2@1 y3@2 y4@3 y5@4 y6@5 y7@6 y8@7;
      ! i-q@0;    ! yields LCGA model, but commented out here ;

OUTPUT: RESIDUAL stand  tech8;
PLOT: Type is plot1 plot2 plot3;  series = y1-y8 (s);
```

Growth mixture model: Programmsyntax in M*plus*

```
TITLE: course of IDS, linear and quadratic, 4 classes, with random effects;
DATA: FILE IS ids.dat;

VARIABLE: NAMES ARE STUDIE THGRUPPE PATNR BDIT1SUM BDIT3SUM
          HA21SUT3 y1 y2 y3 y4 y5 y6 y7 y8 ;
          USEVAR = y1-y8;
          IDVARIABLE=patnr;
          MISSING ARE ALL (999);
          CLASSES=C(4);

 ANALYSIS: TYPE=Mixture missing;
           estimator=ML;
           STARTS = 500 50;
  MODEL:
   %OVERALL%
        i s q | y1@0 y2@1 y3@2 y4@3 y5@4 y6@5 y7@6 y8@7;

OUTPUT: RESIDUAL stand  tech8;
PLOT: Type is plot1 plot2 plot3;  series = y1-y8 (s);
SAVEDATA:  FILE is c:\mplus\idst8c4_cprob; save=cprob;
```

How many classes?

„Problem“:

number of classes is not a model parameter.

Several conceptual approaches are helpful / necessary

- fit criteria
- information criteria (IC), e.g. Bayesian IC (BIC), AIC
- parsimony, theoretical justification, clinical interpretability
- high membership probabilities, classes not too small
- (bootstrap) likelihood ratio tests

How many classes?
Fit and information criteria for GMM with 3 – 5 classes

number of classes	Log Lik.	# of parameters	AIC	BIC	ssaBIC	entropy
3	-7441.8	25	14933.7	15026.6	14947.3	.804
4	-7424.8	29	14907.6	**15015.4**	14923.4	.786
5	-7415.4	33	**14896.8**	15019.5	**14914.8**	.788

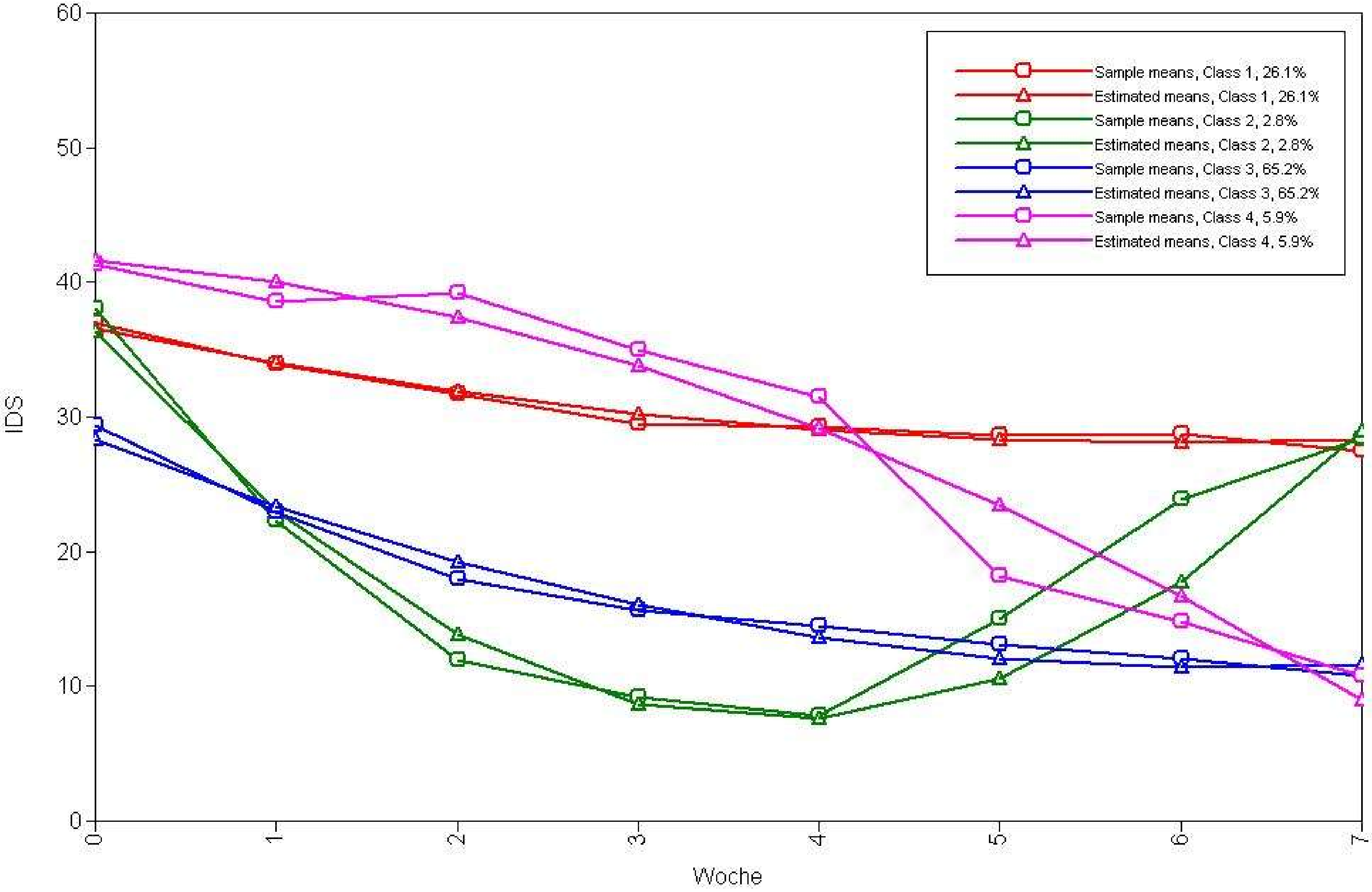
Deciding on the number of classes…

WHAT TO KEEP TRACK OF…

- Log Likelihood value
- # of free parameters
- Model chi-square, df, p-value
- BIC
- Entropy
- LMR-LRT p-value (Tech11)
- BLRT p-value (Tech14)
- Smallest class count and proportion
- Lack of convergence
- Non-replicated Log Likelihood
- Non-positive definite information matrix, etc.

from: Masyn & Muthen

GMM, solution with 4 classes (M*plus*)



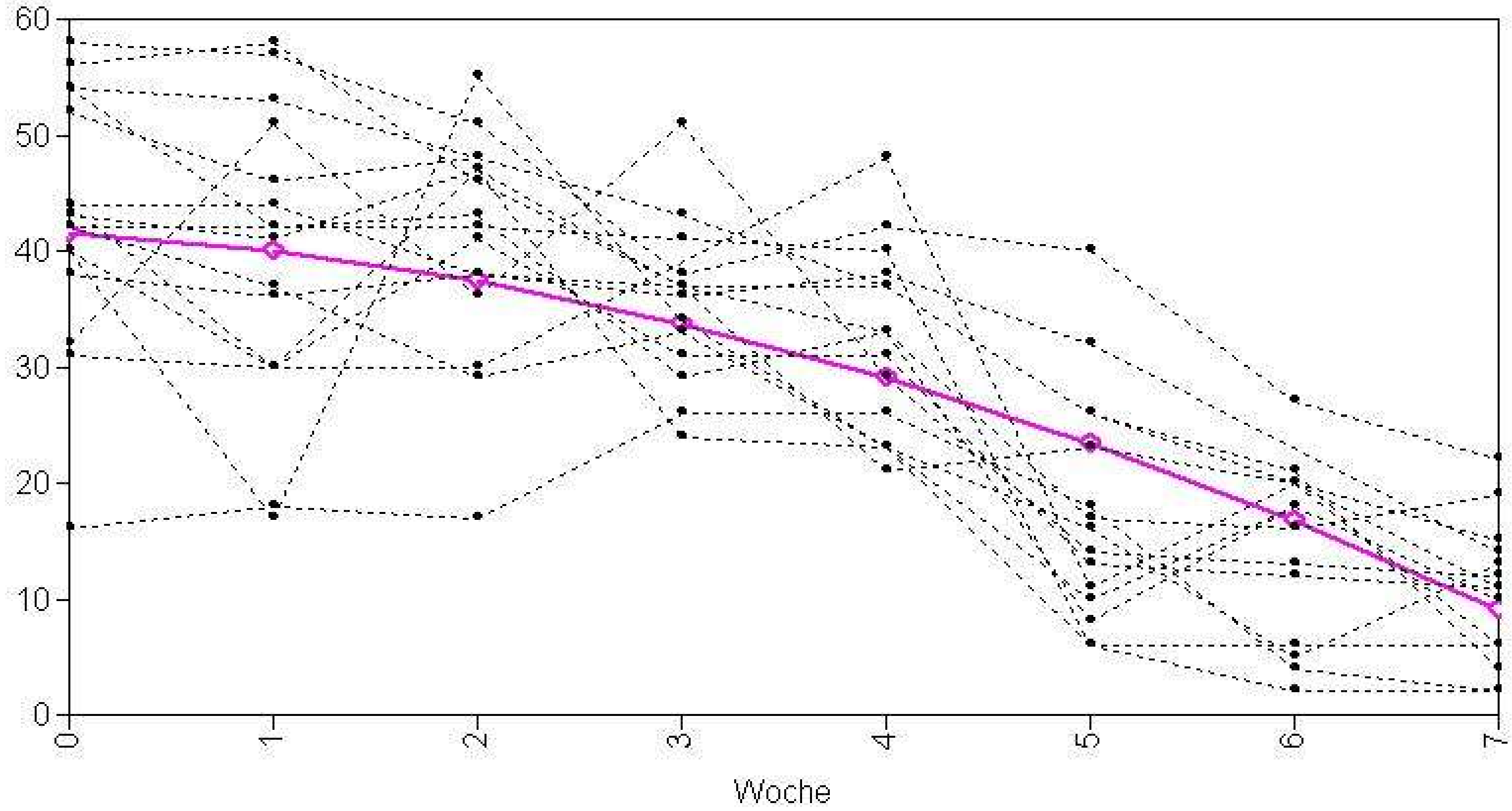
GMM, solution with 4 classes (M*plus*)
(class membership prob. and class attribution)

The SAS System

membership probabilities and modal class

Obs	patnr	prob1	prob2	prob3	prob4	class
1	1101	0.001	0.000	0.446	0.552	4
2	1102	0.006	0.000	0.991	0.002	3
3	1103	0.000	0.000	0.871	0.129	3
4	1104	0.000	0.000	0.994	0.006	3
5	1105	0.000	0.000	0.997	0.003	3
6	1106	0.049	0.000	0.624	0.327	3
7	1107	0.006	0.029	0.856	0.109	3
8	1108	0.021	0.000	0.974	0.005	3
9	1110	0.000	0.000	0.050	0.950	4
10	1113	0.000	0.004	0.973	0.022	3
11	1114	0.000	0.000	0.996	0.004	3
12	1115	0.000	0.974	0.022	0.004	2
13	1116	0.000	0.000	1.000	0.000	3
14	1118	0.026	0.000	0.540	0.434	3
15	1119	0.000	0.000	0.999	0.001	3
...	...	...				

Observed individual courses and estimated trajectory for class 4 („delayed response“)



Keller, F. & Hautzinger, M. (2007). Klassifikation von Verlaufskurven in der Depressionsbehandlung: Ein methodischer Beitrag. *Zeitschrift für Klinische Psychologie und Psychotherapie, 36,* 83-92

„Reality“ of latent classes

LATENT CLASS IDENTITY CRISIS

- Who are we?
- Are we “real”?
- How many are we?
- What defines us?
- What predicts us?
- What do we predict?

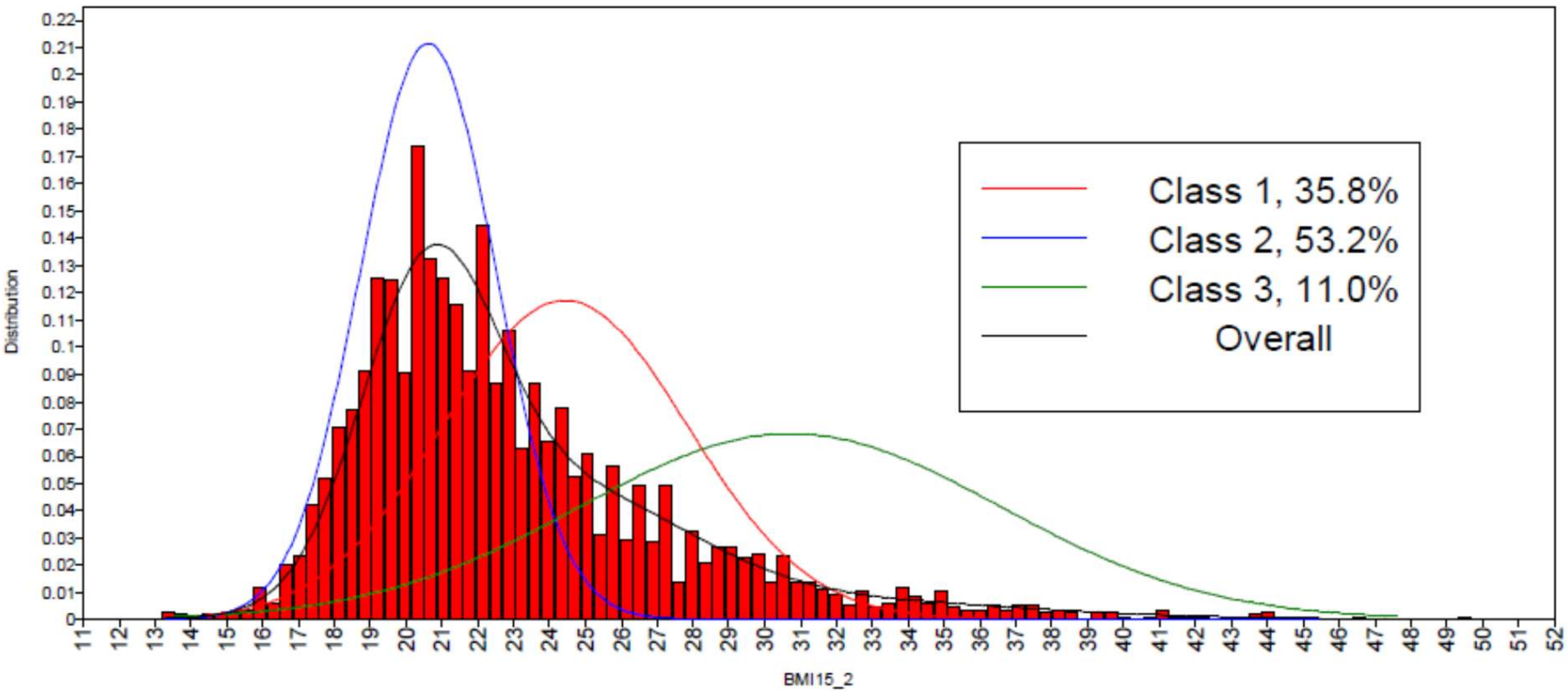
REITOX Academy, EMCDDA © Masyn & Muthén (13-Mar-2009) 60

Mixture of several classes or „simply“ non-normal distribution?

- Discussion on overextraction of latent trajectory classes in 2003 in *Psychological Methods* (Bauer & Curran; Rindskopf; Muthen; and others; reply by Bauer/Curran)

Explaining example: BMI of 15 year old boys

(from a presentation of Muthen, 2014, to PSMG (see at www.statmodel.com))



Kinder- und Jugend-
psychiatrie / Psychotherapie
Universitätsklinikum Ulm

**Klinik für Kinder- und Jugendpsychiatrie /
Psychotherapie des Universitätsklinikums Ulm**

Steinhövelstraße 5
89075 Ulm

www.uniklinik-ulm.de/kjpp

Ärztlicher Direktor: Prof. Dr. Jörg M. Fegert