Multilevel Models

Philippe Rast

PSC 204B
UC Davis, Winter 2018

Topics

Single-level Regression:

Week 1 Linear Regression (G&H: 3,4)
Week 2 Multiple Regression
Week 3 Violation of Assumptions
Week 4 Logistic Regression and GLM (G&H: 5, 6)
Week 5 Over-fitting, Information Criteria and Model comparison (McE: 6)
Week 6 Regression inference via simulations (G&H: 7–10)

Multilevel Regression:

Week 7 Multilevel Linear Models (G&H: 11–13)
Week 8 Multilevel Generalized Models (G&H: 14, 15)
Week 9 Bayesian Inference (G&H: 18 / McE: 1, 2, 3)
Week 10 Fitting Models in Stan and brms (G&H: 16, 17 / McE: 11)

Overview

1. Linear Mixed Models
   - Intraclass Correlation
   - Fixed Effects
   - Random Effects
   - Explanatory Variables
   - Assessment of Model Fit
   - Guidelines for Model Building

Overview

- Hierarchical data
- Merits of longitudinal data
- An example: The BOLSA intelligence data
- Linear mixed models: Fixed effects
- Linear mixed models: Random effects
- Linear mixed models: Covariances between random effects
- Linear mixed models: Explanatory variables
- Linear mixed models: Assessment of model fit
- General guidelines for linear mixed model building

Hierarchical Data

- Participants' data often come in natural groups or clusters, e.g.,
  - Participants belong to the same family
  - Participants live in the same residential home
  - Participants live in the same neighborhood
  - Participants work in the same organization
  - Participants visit the same classes or courses
  - etc.
  - Participants have been assessed several times (repeated measures)

Hierarchical Data

- This grouping or clustering has the effect that participants within groups or clusters are more similar to each other, because
  - Participants share the same upbringing
  - Participants share the same nurses, nutrition, etc.
  - Participants share the same environment
  - Participants share the same organizational climate
  - Participants share the same teachers
  - etc.
  - Or: Participants are the same (repeated measures)

Hierarchical Data

- Grouped or clustered data are called
    - *hierarchical*, because individual observations are nested into higher-order units
    - *multilevel*, because one can distinguish different data levels, e.g., individual versus group level
- There may be more than two levels of hierarchy, e.g.,
    - Participants (level 1) are nested into the classes (level 2), classes are nested into schools (level 3)
    - Or: Observations (level 1) are nested into participants (level 2), participants are nested into families (level 3) (repeated measures)

Hierarchical Data

- Hierarchical or multilevel data lead to dependent observations (in the statistical sense)
- Consequences for analytical approaches ignoring this dependency, e.g., OLS regression:
  - A single observation does not contribute as much information as is assumed, because
  - The assumption of i.i.d. (independent and identically distributed) observations is violated
  - Standard errors of parameters are biased
  - Statistical significance tests will thus lead to the wrong conclusions

Hierarchical Data

- However, the severeness of these consequences depends on the amount of data dependency, because:
  - If data dependency is small, approaches ignoring data dependency will hardly affect results of significance tests
  - If data dependency is large, approaches ignoring data dependency will strongly affect results of significance tests
- How, then, to decide whether the amount of data dependency is ignorable (small) or not (large)?

Hierarchical Data

- The amount of data dependency depends on the proportion of variance between groups or clusters in relation to the total variance, i.e.:

$$\rho_{ic} = \frac{d_{11}}{d_{11} + \sigma^2} \tag{1}$$

  where d_{11} = variance between classes, i.e., differences between classes
  σ^2 = variance within classes, i.e., individual differences

- ρ_{ic} is called the *intraclass correlation coefficient* (ICC, or cluster effect)

Hierarchical Data

- Range of intraclass correlation

$$0 < \rho_{ic} < 1$$

No data dependency Complete data dependency

- Rule of thumb: $\rho_{ic} > .05$ may be considered substantial (for parameter estimation, standard errors, etc.)
- For repeated measures data, ρ_{ic} is virtually always substantial

Hierarchical Data

- Another perspective: Cluster effect

$$\mathrm{CE} = 1 + (n_i - 1)\rho_{ic}$$

  where n_i denotes the average cluster size.
- Rule of thumb: $\mathrm{CE} \geq 2$ indicates that the clustering in the data needs to be taken into account during parameter estimation.
  - Example:
    $\rho_{ci} = .15$
    If $n_i = 5$ then $\mathrm{CE} = 1.6$
    If $n_i = 15$ then $\mathrm{CE} = 3.1$

    $\rho_{ci} = .55$
    If $n_i = 5$ then $\mathrm{CE} = 3.2$
    If $n_i = 15$ then $\mathrm{CE} = 8.7$

Hierarchical Data

- Advantages of *multilevel models*
  - They solve the “unit of analysis”-problem

Example: In comparing residential homes, observations are based on individuals (level 1), but interest may lie in differences between residential homes (level 2)

  - Allow for simultaneous analyses at different levels

Example: From a behavioral genetics perspective, it seems interesting to study both individual (level 1) and family (level 2) differences in intelligence

  - Inclusion of explanatory variables at different levels

Example: Language achievement may be explained by both the com-position of classes (e.g., number of students, percentage foreigners at level 2) and individual variables (e.g., foreigner?)

  - Regularization via shrinkage
    - ▷ Property of estimation which takes into account reliability of estimates within a given group/individual

Hierarchical Data

Some Thoughts on Longitudinal Models

- In what follows, focus will be on *linear mixed models for repeated measures data*, but
  - With the according modifications, similar models may be applied to hierarchical data
  - Principles of estimation are the same, but error variances and covariances may be modeled in a different way

- Most examples are from longitudinal data
  - ▷ Some thoughts about longitudinal modeling

Merits of Longitudinal Data

- Effectiveness in studying change

Example: Development of intelligence

- Cross-sectional model:

$$y_i = \beta_C x_i + \epsilon_i$$

where y_i is the intelligence score of individual i
x_i is the age of individual i
β_C is the regression of intelligence on age ϵ_i is a residual

- Then, β_C represents the difference in average y across two sub-populations which differ by 1 year of age

Merits of Longitudinal Data

- *Longitudinal model:*

$$y_{ij} = \beta_C x_{i1} + \beta_L (x_{ij} - x_{i1}) + \epsilon_{ij}$$

where y_{ij} is the intelligence score of individual *i* at age j
x_{i1} is the age of individual *i* at first measurement occasion
β_C is the cross-sectional regression of intelligence on age
x_{ij} is the age of individual *i* at *j*th measurement occasion
β_L is the regression of intelligence on age changes
ϵ_{ij} is a residual

- Then, β_L represents the average longitudinal change in *y* across individuals who changed by 1 year of age

Merits of Longitudinal Data

- To estimate how individuals change with time from cross-sectional data, we must make the strong assumption that $\beta_C = \beta_L$
- With longitudinal data, this strong assumption is unnecessary, because both β_C and β_L can be estimated
- Even if $\beta_C = \beta_L$, longitudinal data tend to be statistically more powerful
    - The basis of inference about β_C is a comparison of individuals with a particular value of age to others with a different age.
    - By contrast, β_L is estimated by comparing a person's intelligence at two (or more) times.
    - Longitudinally, each person can be thought of as serving as her own control, thus canceling out the influence of unmeasured characteristics in estimating β_L, whereas they tend to obscure the estimation of β_C

Merits of Longitudinal Data

- Distinguish the degree of variation in y across time for one person from the variation in y among people
- With cross-sectional data, the estimate of one person must draw on upon data from other persons to over-come measurement error, which, however, ignores individual differences
- With repeated measures, strength can be borrowed from observations across time for one person and for other persons
- If there are little individual differences, individual estimates may also rely on data from others. Else, we might prefer to use only data for a specific person.

Merits of Longitudinal Data

- In practice, longitudinal data are oftentimes highly unbalanced, because
  - An equal number of measurements is not available for all subjects (drop-out, attrition)
  - Measurements are not taken at the same, fixed time points
  - Traditional analysis approaches (ANOVA) rely on balanced data, which leads to a great loss of information
  - As we will see, the same is not true for mixed models, which can use all available information (under special assumptions)

Longitudinal Hierarchical Models

Example: The BOLSA Data

Bonn Longitudinal Study on Aging (BOLSA)

- Seven measurement occasions with decreasing sample size

Table: BOLSA Data

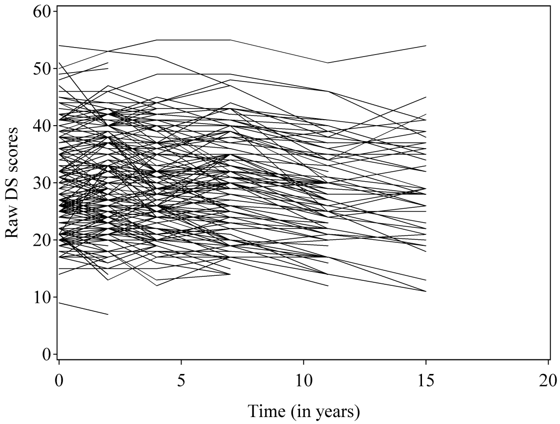
	1965	1967	1969	1972	1976	1980	1984
N=	221	188	158	127	80	46	30

- Two cohorts:
  - 113 individuals aged 63.3 years at T1
  - 108 individuals aged 72.4 years at T1
- Focus here: WAIS subtests Digit Symbol (DS; only at T1 – T6) and Block Design (BD)

Longitudinal Hierarchical Models

Example: The BOLSA Data

- Individual trajectories in **DS** (“Spaghetti Plot”)

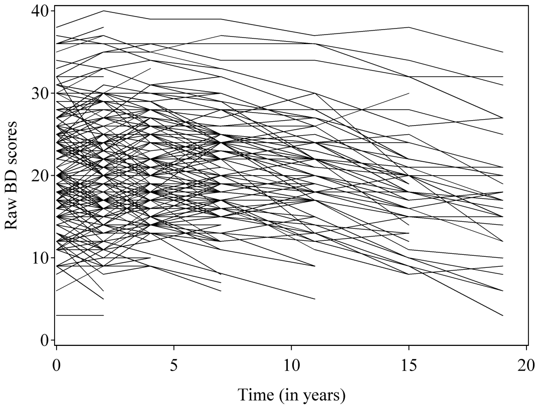


- Considerable heterogeneity (both in initial level and slope)
- Trajectories appear to be roughly linear
- On average, DS performance seems to decline slightly

Longitudinal Hierarchical Models

Example: The BOLSA Data

- Individual trajectories in **BD** ("Spaghetti Plot")



- Considerable heterogeneity (both in initial level and slope)
- Trajectories appear to be roughly linear
- On average, BD performance seems to decline slightly

Longitudinal Hierarchical Models

Example: The BOLSA Data

- Focus is on change in DS for those subjects with complete data for the first five measurement occasions ($N = 81$, longitudinal time period: 11 years)
- Note that the outcome variable is continuous (For ordinal and categorical outcome variables, other, more complicated models are needed)
- Note that we want to model change in one variable (univariate case), not in several variables simultaneously (multivariate case)
- Note that we only use complete cases in the next examples

Longitudinal Hierarchical Models

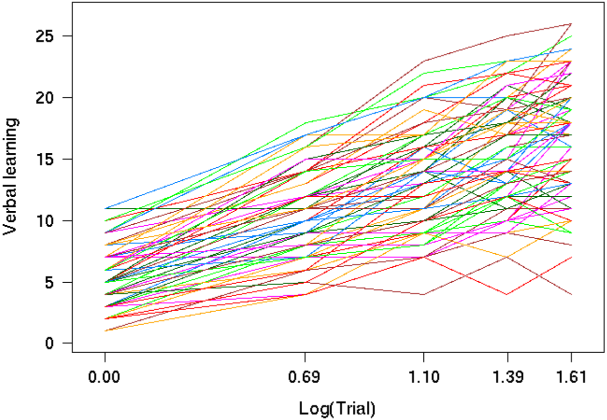
Example: The ZULU Data

- Zurich Longitudinal Study on Cognitive Aging (ZULU)
- $N = 364$ (46% female) ranging from 65 to 80 years ($M = 72.98$, $SD = 4.42$)
- At the first measurement occasion in 2005 verbal learning was assessed using 5 trials and 27 words. Verbal learning was operationalized as the recall performance after each learning trial.
- For this example wie will use 70 randomly drawn participants.

Longitudinal Hierarchical Models

Example: The ZULU Data

- Raw scores from 5 recall trials: Trajectories tend to be non-linear
- The learning trajectories are linearized by log-transforming the Trial variable



Longitudinal Hierarchical Models

Example: The ZULU Data

- Model level and slope and individual differences in level and slope, across five measurement occasions.
- The trial number has been log-transformed in order to linearize the trajectories
  - ▷ We could also model truly non-linear mixed effects models

Longitudinal Hierarchical Models

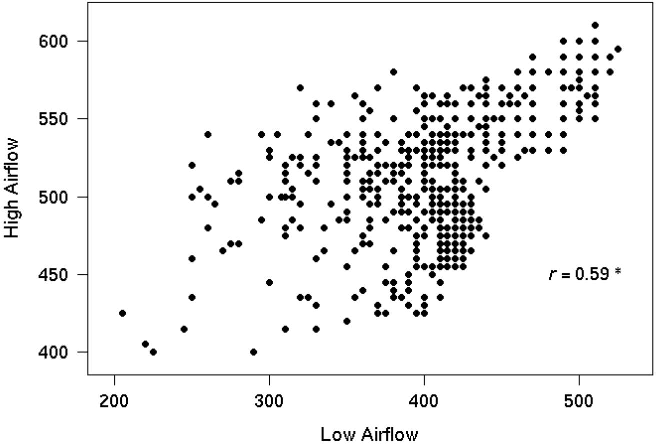
Example: The Asthma Data

- 80 participants with slight symptoms of asthma.
- At maximally 16 days the upper and lower range of the peak flow were measured.
- The peak flow indicates a participants' maximum ability to exhale. Peak flow readings are higher when participants are well, and lower when the airways are constricted.
- The best and worst of three readings are used as the recorded value of the upper and lower Peak Expiratory Flow Rate.

Longitudinal Hierarchical Models

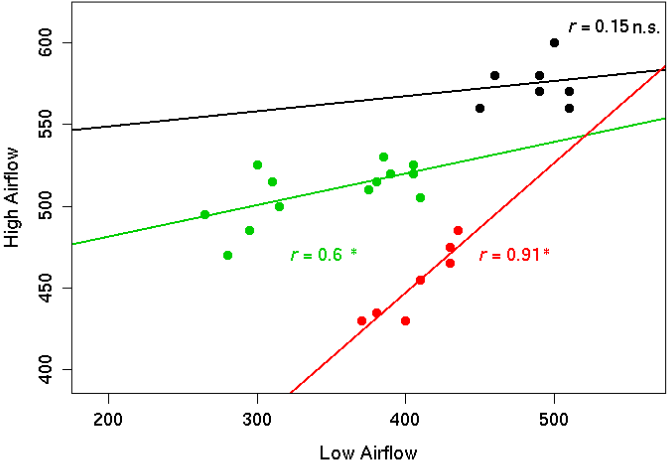
Example: The Asthma Data

- Correlation between high and low airflow (ignoring data dependency)



Example: The Asthma Data

- Considerable individual differences between participants



Linear Mixed Models

- What are linear *mixed* models?

Synonyms: *hierarchical* linear models
linear *multilevel* models
linear *variance components* models

- Mixed models contain both *fixed* and *random effects*
  - *mixed* models
- Hierarchical and multilevel
- Variance components refers to the fact that residual variance is split in systematic and error variance, i.e., different components

Linear Mixed Models

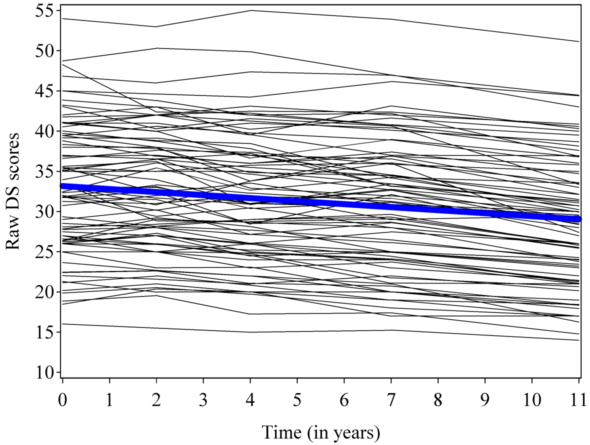
- Why are these models linear?
  - Because predictor variables enter linearly
  - Thus, the model is linear in its parameter estimates, i.e., the associations between predictor variables and the outcome variable are considered to be linear
  - Note: Like in OLS regression, one may still model polynomial associations (e.g., quadratic, cubic) by raising the predictor variable to the according power
- Can be expanded to intrinsically nonlinear models

Linear Mixed Models: Fixed Effects

- What are fixed effects?
- *Fixed effects* describe average effects, e.g., the average initial status and the average change across time
- Fixed effects are called fixed because it is assumed that they reflect what is happening in the population, i.e., average effects for an "average" person
- Thus, without having any further information about a person – apart from that she belongs to the population captured by the sample – the fixed effects would be the best "blind" guess in describing such a person's longitudinal trajectory

Linear Mixed Models: Fixed Effects

- Individual trajectories in DS across 11 years ($N = 81$)



Linear mixed model mean trajectory:
$DS =$
$33.33 - 0.39 \times Time$

Linear Mixed Models: Fixed Effects

- Statistical analyses:
  - OLS Regression (ignoring data dependency):
    $DS = 33.33 - 0.39 \times Time$
  - Mixed model (taking data dependency into account):
    $DS = 33.33 - 0.39 \times Time$
  - In this example, (fixed) parameter estimates are the same, why?
  - In general, if and only if the data are balanced, estimation of (fixed) parameters using OLS regression or mixed models will give identical results

Linear Mixed Models: Fixed Effects

- Fixed parameter estimates for the BOLSA data are the same, but what about the standard errors and statistical significance?

	OLS	Mixed
Intercept	33.33	33.33
$S.E.$	0.66	0.89
t	50.51	37.20
Slope	-0.39	-0.39
$S.E.$	0.11	0.03
t	-3.54	-12.48

1. Considerable differences in standard errors
2. Considerable differences in t values

Linear Mixed Models: Fixed Effects

- The same holds for the ZULU data:

	OLS	Mixed
Intercept	5.29	5.29
$S.E.$	0.42	0.31
t	12.62	17.31
Slope	7.17	7.17
$S.E.$	0.38	0.28
t	19.00	25.55

1. Considerable differences in standard errors
2. Considerable differences in t values

Linear Mixed Models: Fixed Effects

- Why are the standard errors different and, hence, the results of statistical significance tests?
  - Like with repeated measures ANOVA, from longitudinal data strength may be borrowed because individuals function, in a sense, as their own controls
  - Hence, like in repeated measures ANOVA, in mixed models the mean trajectory reflects the mean of individual trajectories, and not of unrelated observations (data points)
  - However, whereas in repeated measures ANOVA a saturated model is estimated, i.e., a parameter for each mean, in linear mixed models mean changes are related linearly to time (more parsimonious)

Linear Mixed Models: Random Effects

- Individual trajectories may be different from the mean longitudinal trajectory
- These individual departures from the mean trajectory are called *random effects*
- Important: The assumption is that individual trajectories are of the same functional form, e.g., linear, as the mean trajectory
- Thus, if, e.g., the mean trajectory is modeled using two parameters (initial level, linear slope), random effects may exist in initial level and linear slope
- Whereas fixed effects are denoted using Greek letters, random effects are denoted using Latin letters

Linear Mixed Models: Random Effects

- Whereas the mean trajectory is modeled at level 2 (the individual or between-person level), individual trajectories are modeled at level 1 (the observational or within-person level)
- Thus, for a model of linear change, we have

  Level 1: $y_{ij} = b_{i0} + b_{i1}t + e_{ij}$

  - y_{ij} is the outcome variable of person i at measurement occasion j
  - b_{i0} is the individual-specific or random intercept of person i
  - b_{i1} is the individual-specific or random slope of person i
  - t is the time elapsed since first measurement occasion

Linear Mixed Models: Random Effects

- At level 2, we have two models

  Level 2: $b_{i0} = \beta_0 + u_{i0}$

  $$b_{i1} = \beta_1 + u_{i1}$$

  - β_0 is the fixed intercept
  - u_{i0} is the individual departure from the fixed *intercept* of person i
  - β_1 is the fixed slope
  - u_{i1} is the individual departure from the fixed *slope* of person i

Linear Mixed Models: Random Effects

- The level 1 and level 2 models may be combined to a complete model, namely
  $y_{ij} = \beta_0 + u_{i0} + (\beta_1 + u_{i1})t + e_{ij}$

- More general mixed effects notation (Laird-Ware form):

$$
\begin{aligned}
\boldsymbol{y}_i &= \boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{Z}_i \boldsymbol{u}_i + \boldsymbol{\epsilon}_i \\
\boldsymbol{u}_i &\sim MVN(\mathbf{0}, \boldsymbol{\psi}) \\
\boldsymbol{\epsilon}_i &\sim MVN(\mathbf{0}, \sigma^2 \boldsymbol{\Omega})
\end{aligned}
$$

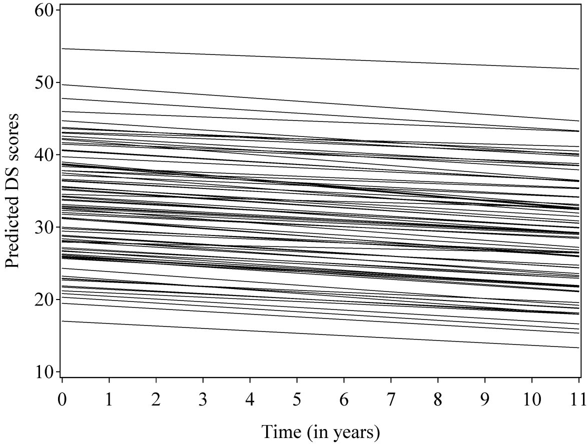
$\sigma^2\boldsymbol{\Omega}$ is also referred to as the **R**-Matrix.
$\boldsymbol{\Omega}$ may take different forms

Linear Mixed Models: Random Effects

- Of interest are most often not the individual parameters, but rather whether
  - Random effects show statistically significant variance across persons, because
  - Significant variance implies that, e.g., persons differ reliably in initial level and in the amount of linear slope, i.e., develop differentially

Linear Mixed Models: Random Effects

- Plot of model-based predicted DS Scores (N = 81)



- Individual differences in initial level
- Individual differences in linear slope

Linear Mixed Models: Random Effects

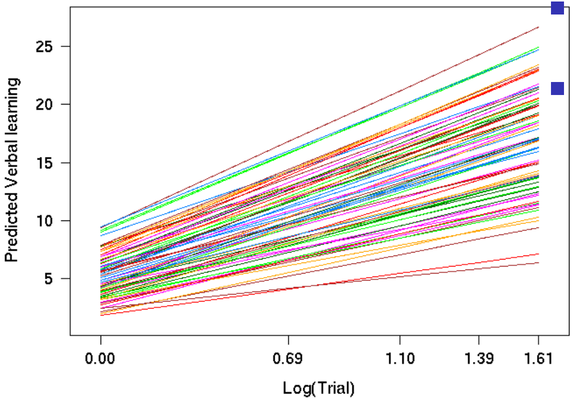
- Results of random effects estimation

Parameter	Estimate	S.E.	Z	*p*
Level Variance	62.75	10.22	6.14	$<.01$
Slope Variance	0.04	0.01	3.13	$<.01$
Error Variance	2.88	0.26	10.95	$<.01$

- Both the level and especially the slope variance are statistically significant
- This implies that there are reliable individual differences in initial DS level and the amount of linear change in DS across time

Linear Mixed Models: Random Effects

- Plot of model-based predicted verbal learning scores.



- Individual differences in initial level
- Individual differences in linear slope

Linear Mixed Models: Random Effects

- Results of random effects estimation

Parameter	Estimate	S.E.	Z	*p*
Level Variance	4.56	1.13	4.03	$<.01$
Slope Variance	3.92	0.95	4.11	$<.01$
Error Variance	2.59	0.25	10.25	$<.01$

- Both the level and the slope variance are statistically significant
- This implies that there are reliable individual differences in initial verbal learning level and the amount of linear change in verbal learning across time

Linear Mixed Models: Random Effects

- Note that with a repeated measures ANOVA, we couldn't have reached the same conclusion, since
  - Whereas the random effect in initial level is somewhat trivial (nobody would have expected that every person starts out from the same performance level),
  - The random effect of linear change (slope) is more interesting, because it corresponds to an interaction between individual and time
  - Why can't we find such an interaction with ANOVA?
  - Because ANOVAs are saturated models (for every of the 5 measurements a separate parameter is estimated, i.e., no degrees of freedom are left)
  - In our model, we used (only) two parameters to describe change across 5 occasions , i.e., there are some *df*'s left

Linear Mixed Models: Random Effects

- In this respect, linear mixed models are more similar to regression models than to ANOVAs, because
  - The goal is to find an adequate, but most parsimonious (with respect to the number of parameters) description of an association (here: between the outcome variable and time)
  - Had we needed more parameters, we would have used up more degrees of freedom, until we would have reached a saturated model (with 5 parameters) as well
  - As with every model building process, the goal is adequacy and parsimony

Linear Mixed Models: Random Effects

- Hence, there is an important rule for linear mixed models:
- In order to estimate random effects, the number of parameters must be smaller than the number of measurement occasions
- Thus, with only two measurement occasions, we can estimate a random initial level, but no random effect for change, although for both we could estimate fixed effects
- Clearly, the hypothesis of, e.g., linear change can be tested more strictly the more measurement occasions are available

Covariances between random effects

- Random effects are, in essence, individual differences variables
- Thus, like with all individual differences variables, we might be interested in the association between random effects, in this case: Between initial level and change
- Here, this would answer the question whether those persons starting out at a high performance level change more or, in turn, less than those who start out with a lower performance level
- For the BOLSA DS data, a correlation of $-.17$ (not significant) between initial level and linear change is estimated. For the ZULU data, the same correlation of $.11$ is significant.

Covariances between random effects

- A "problem" in longitudinal linear mixed models is that the covariance between level and slope depends on how level is defined
- Thus, the covariance is different if level is defined as initial level, intermediate level, or end level
- This makes, after thinking about it, perfect sense, because these are different models
  - Watch out when interpreting the covariance
- Another, real problem is that the covariance may be influenced by ceiling or floor effects, because, e.g., it is hard to measure decline in those who already start out with a very low performance level

Explanatory Variables

- Once we have found significant fixed and/or random effects, the next logical step is to try to "explain" these effects
- For the BOLSA DS data, we might ask, e.g., whether cross-sectional age differences can explain fixed and random effects in initial level and slope
- Note that continuous explanatory variables have to be grand mean centered in order to keep the fixed effects interpretable
- But since age cohort is a dichtomy ($0 =$ younger cohort, $1 =$ older cohort), 0 is a meaningful value, since then the fixed effects represent effects for the younger cohort

Explanatory Variables

- Results with cohort as an explanatory variable

	Without cohort	With cohort
Level	33.33*	34.98*
Slope	-0.39*	-0.33*
Level$\times$Cohort	–	-3.88*
Slope$\times$Cohort	–	-0.15*
Var(Level)	62.75*	60.22
Var(Slope)	0.04*	0.04*
Corr(LS)	-0.13	-0.23

Assessment of Model Fit

- Linear mixed modeling involves trying several different models for the data at hand an then decide which model fits the data best (important: also under theoretical considerations)
- For linear mixed models, two models may be compared
  - According to their -2$\times$log-likelihood
  - According to Akaikes Information Criterion (AIC)
- Both indices are **relative**, not absolute, i.e., they make sense only in the comparison of two (or more) models
- For both indices, lower values indicate better model fit AIC penalizes parametrization and encourages parsimony

Assessment of Model Fit

- $-2\times$log-likelihood is useful in two models that are "nested", i.e., which are the same with respect to the involved effects, but different in the number of estimated parameters (e.g., by estimating one covariance between random effects more). The difference in $-2\times$log-likelihood may then be tested for statistical significance using a χ^2-test
- AIC is useful in comparing any two models, but there are no clear-cut thresholds in deciding which model fits better
- In general, compared to SEM, for linear mixed models emphasis on absolute model fit is not that high, which is also reflected by the small number of available fit indices (compared to SEM)

General Guidelines for Model Building

- There are three steps in building a linear mixed model

  1. Selection of a (preliminary) mean structure
  2. Selection of a (preliminary) random effect structure
  3. Selection of a residual covariance structure

General Guidelines for Model Building

- Selection of a (preliminary) mean structure
  - In a first step, an overelaborated model for the mean response profile may be used
  - This coincides with the traditional ANOVA
  - However, like with the traditional ANOVA, the concept of a saturated model breaks down when data are unbalanced
  - As an alternative, a plot of a smoothed average trend of individual profiles helps to select a candidate mean structure (ignoring data dependency)
  - Some authors recommend using general additive models.
  - Generally, good idea to plot data

General Guidelines for Model Building

- Selection of a (preliminary) random effect structure
  - Check for variability changes over time.
  - In general, the inclusion of too many random effects rather than omitting some appears to be a favorable strategy, because
  - This ensures that the remaining variability is not due to any missing random effects
  - Afterwards, statistically non-significant random effects may be skipped in the process of model reduction
  - But, some research shows that even when non-significant random effects are omitted, fixed effects tend to be biased (Type I error increase)

General Guidelines for Model Building

- Selection of a residual covariance structure
  - Conditional on the selected set of random effects, the specification of the error variance matrix Σ_i is needed
  - However, there are no simple techniques available to compare different error variance structures
  - One strategy is to fit a series of models with the same mean and random effects structure, but with different error variance structures
  - These different models may then be compared using likelihood-based criteria
  - However, unless one is especially interested in the error variance structures, it is usually sufficient to compare a few models

Example: Cebu Data

- The Cebu Longitudinal Health and Nutrition Survey
- City in the Philippines
- "onceptualized as an interdisciplinary study of infant-feeding patterns, particularly the overall sequencing of feeding events (milks and complementary foods)"
- cf. Adair et al. (2011)