Data analysis using R – Statistical Epidemiology

12

Linear mixed models II

Stefan Konigorski

Health Intervention Analytics Group, HPI

February 3, 2022

HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Content

Block	Class	Content	Date
R, Data manipulation, Descriptives	1	Overview & Introduction to R and data analysis	2021.10.28
	2	First steps in data analysis using R	2021.11.04
	3	Second steps in data analysis using R	2021.11.11
Epidemiology & Statistics: concepts	4	Epidemiological study designs	2021.11.18
	5	Estimation	2021.11.25
	6	Hypothesis testing & study planning	2021.12.02
	7	Missing data	2021.12.09
Data analysis w/ regression models	8	Linear regression I	2021.12.16
	9	Linear regression II	2022.01.13
	10	Regression models for binary and count data	2022.01.20
	11	Analysis of variance & Linear mixed models I	2022.01.27
	12	Linear mixed models II & Meta analysis	2022.02.03
	13	Survival analysis	2022.02.10
	14	Causal inference & Data analysis challenge	2022.02.17

(see full schedule online)

Review of last week

- ANOVA: variance decomposition and more.
- Introduction to linear mixed models: understand the idea behind fixed and random effects.

Learning objectives of today

- Linear mixed models for hierarchical data.
- Outline of linear mixed models for longitudinal data.
- Application of mixed models to meta-analysis.

Linear mixed models for hierarchical data

What are fixed and random effects?

What are fixed effects (of predictors)?

- Fixed effect = parameter (β or βX) that refers to a variable with fixed values that are of interest.
- Population-average effects
- Example: Sex with values "male", "female".

What are random effects?

- Random effect = parameter that refers to a variable denoting clusters (=groups) or individuals that have been drawn randomly and whose values are not of interest per se.
- Subject-specific effects
- Examples: dataset with students from 20 randomly selected schools $\longrightarrow$ school = random effect.

Example: Mathematical Achievement dataset

Dataset `MathAchieve`

- Available in `nlme` package.
- Dataset on "Mathematics achievement scores" (variable `MathAch`) of 7185 students with further variables `School`, `Minority` (yes/no), Sex, `SES` (socioeconomic status), `MEANSES` (mean SES of the school in which the students are).

Main research question

- Is the SES associated with the math scores?

Fixed-effects (only) model

Model

- Predict the grade of each child $i = 1 \ldots n$ by the general mean, SES, sex and minority:

$$\text{Math}_i = \beta_0 + \beta_1 SES_i + \beta_2 Sex_i + \beta_3 Minority_i + \varepsilon_i$$

- Multiple lineare regression with $\beta_0, \beta_1, \beta_2, \beta_3$ as fixed effects.

Which assumptions does this model contain?

- The math grade only depends on the SES, sex and minority.

in R

- `lm(MathAch ~ SES + Sex + Minority, data = MathAchieve)`

$\longrightarrow$ Conclusion: very little variance explained!

How can school be considered in the analysis?

Last week

- Include dummy variable for every school in regression/ANOVA
- Include school as continuous variable in the regression.
- Stratified analysis, separately for each school.
- Analysis not on individual but school level.

How can school be considered in the analysis?

Last week

- Include dummy variable for every school in regression/ANOVA
- Include school as continuous variable in the regression.
- Stratified analysis, separately for each school.
- Analysis not on individual but school level.

Today

- School as random effect in a mixed model.
- (a) ... in mixed model with random intercept.
- (b) ... in mixed model with random intercept and random slope.

Random intercept model

Model

$$\text{Math}_{ij} = (\beta_0 + \gamma_{0j}) + \beta_1 SES_{ij} + \beta_2 Sex_{ij} + \beta_3 Minority_{ij} + \varepsilon_{ij}$$

of student i in school j with $\gamma_0 \sim N(0, \sigma_1^2)$, $\varepsilon \sim N(0, \sigma_2^2)$, where γ_0 and ε are independent.

Assumption

Association between SES and math grade is the same in all schools (same slope β_1) except for random constant (γ_0).

Random intercept model

in R

- `nlme::lme(MathAch ~ SES + Sex + Minority, random = ~1|School, data = MathAchieve)`
- `lme4::lmer(MathAch ~ SES + Sex + Minority + (1|School), data = MathAchieve)`

Random intercept, random slope model

Model

$$\text{Math}_{ij} = (\beta_0 + \gamma_{0j}) + (\beta_1 + \gamma_{1j})SES_{ij} + \beta_2 Sex_{ij} + \beta_3 Minority_{ij} + \varepsilon_{ij}$$

of student i in school j with $\gamma_0 \sim N(0, \sigma_1^2)$, $\gamma_1 \sim N(0, \sigma_2^2)$, $\varepsilon \sim N(0, \sigma_3^2)$, where γ_0 and ε as well as γ_1 and ε are independent, γ_0 and γ_1 may correlate.

Random intercept, random slope model

Model

$$\text{Math}_{ij} = (\beta_0 + \gamma_{0j}) + (\beta_1 + \gamma_{1j})SES_{ij} + \beta_2 Sex_{ij} + \beta_3 Minority_{ij} + \varepsilon_{ij}$$

of student i in school j with $\gamma_0 \sim N(0, \sigma_1^2)$, $\gamma_1 \sim N(0, \sigma_2^2)$, $\varepsilon \sim N(0, \sigma_3^2)$, where γ_0 and ε as well as γ_1 and ε are independent, γ_0 and γ_1 may correlate.

Assumption

Association between SES and math grade varies by a random constant (γ_0) and random factor (γ_1) between the schools.

Random intercept, random slope model

Model

$$\text{Math}_{ij} = (\beta_0 + \gamma_{0j}) + (\beta_1 + \gamma_{1j})SES_{ij} + \beta_2 Sex_{ij} + \beta_3 Minority_{ij} + \varepsilon_{ij}$$

of student i in school j with $\gamma_0 \sim N(0, \sigma_1^2)$, $\gamma_1 \sim N(0, \sigma_2^2)$, $\varepsilon \sim N(0, \sigma_3^2)$, where γ_0 and ε as well as γ_1 and ε are independent, γ_0 and γ_1 may correlate.

in R

- `nlme::lme(MathAch ~ SES + Sex + Minority, random = ~SES|School, data = MathAchieve)`
- `lme4::lmer(MathAch ~ SES + Sex + Minority + (SES|School), data = MathAchieve)`

What's the use of these (complicated) models?

- The estimates of the fixed effects and their standard errors consider the random effects i.e. the random variation between schools.
- In detail 1: the estimated fixed effect is a weighted mean of the standard regression coefficient estimate and the random effect.
- In detail 2: If observations are highly correlated, this reduces the effective sample size and power - but not as much if you use mixed models compared to dummy variables!
- The not explained variance of the school grades (SS between groups) can be partly explained by the random effects.

Interpretation of the effects in mixed models

Interpretation of the fixed effects

- The fixed effects (i.e. regression coefficients β of the fixed effects) are "conditional effects": conditional on the other predictors (as in standard linear regression) and conditional on the random effects.
- I.e. the regression coefficients describe the mean change in Y for a 1-unit change in X, when holding all other predictors constant and adjusting for the random (normally-distributed) differences between clusters (i.e. for a person in a specific cluster) $\longrightarrow$ "cluster-specific", "subject-specific effects".

Interpretation of the random effects

$\longrightarrow$ see variance components on slide 19.

Interpretation of the effects in mixed models

Visualization of fixed effects

- For a better understanding of the fixed effects and their variation between schools, the school-specific associations between SES and math grade can be inspected.
- The predicted math grade of the students consists of the prediction by the fixed effects plus the school-specific random term.
- The school-specific predictions can be extracted (like in regression) by using the `predict` function (`predict(lme())`).
- A better understanding can often be gained through a visualization, see `R_12a_exercise_LMM.Rmd`.

Context variables

Overview

- Context variables can also be investigated in mixed (hierarchical) models.
- Context variable (= variable on macro level): Variable at group level, here: SES on school level.
- Motivation: The context variables can contain information that variables on individual level don't capture. Eg: SES on school level can be a marker for environmental factors that are relevant for math grade.
- In the interpretation you have to consider this variable level (cf. ecological fallacy): difference in the mean math grade of students on the school level when the mean SES of the school increases by 1 unit.

Context variables

in R

- `nlme::lme(MathAch ~ SES + Sex + Minority + MEANSES, random = ~1|School, data = MathAchieve)`

Variance components and ICC

Variance components

- The variance of Y in a mixed model is equal to the sum of the variances of the random effects.
- eg in random intercept model on slide 10:
  $Var(Y) = Var(\gamma_0) + Var(\varepsilon) = \sigma_1^2 + \sigma_2^2$
- With random slope: $Var(Y) = Var(\gamma_0) + Var(\gamma_1) + Var(\varepsilon)$

Variance components and ICC

ICC = intraclass correlation coefficient

- Proportion of the total variance of Y that can be explained by the variance on the "higher level" (here: school).
- In the random intercept model on slide 10: ICC = $\frac{Var(\gamma_0)}{Var(\gamma_0)+Var(\varepsilon)}$
- ICC is often computed in a random intercept model.
- In model with random slope: ICC is dependent on the value of the respective X variable, i.e. there exist multiple ICCs.

Variance components and ICC

ICC = intraclass correlation coefficient

- Proportion of the total variance of Y that can be explained by the variance on the "higher level" (here: school).
- In the random intercept model on slide 10: ICC = $\frac{Var(\gamma_0)}{Var(\gamma_0)+Var(\varepsilon)}$
- ICC is often computed in a random intercept model.
- In model with random slope: ICC is dependent on the value of the respective X variable, i.e. there exist multiple ICCs.

ICC in R

- Computation directly from the estimated variance of the random effects in the output of e.g. `lme`.
- Or from other functions, zB `psychometric::ICC1.lme()`.
- see `R_12a_exercise_LMM.Rmd`.

Alternative model notation

Intercept-only fixed-effect model

- Predict grade of each student $i = 1 \ldots n$ by the general mean:

$$\text{Math}_i = \mu + \varepsilon_i \text{ or } \text{Math}_i = \beta_0 + \varepsilon_i$$

- $\varepsilon \sim N(0, \sigma^2)$ hence $Y \sim N(\mu, \sigma^2)$.
- μ (or β_0) is fixed effect.

Alternative model notation

Fixed-effect model

- Predict grade of each student $i = 1 \ldots n$ by the general mean and SES:

$$\text{Math}_i = \mu + \gamma_i + \varepsilon_i \text{ or } \text{Math}_i = \beta_0 + \beta_1 SES_i + \varepsilon_i$$

- μ is fixed effect (mean math grade).
- γ is fixed effect (effect of student's SES).

Alternative model notation

Fixed-effect model

- Predict grade of each student $i = 1 \ldots n$ by the general mean and SES:

$$\text{Math}_i = \mu + \gamma_i + \varepsilon_i \text{ or } \text{Math}_i = \beta_0 + \beta_1 SES_i + \varepsilon_i$$

- μ is fixed effect (mean math grade).
- γ is fixed effect (effect of student's SES).

Note: There exist also other equivalent ("hierarchical") notations of mixed models.

Introduction to linear mixed models for longitudinal data

HRS (Health and Retirement Study) dataset

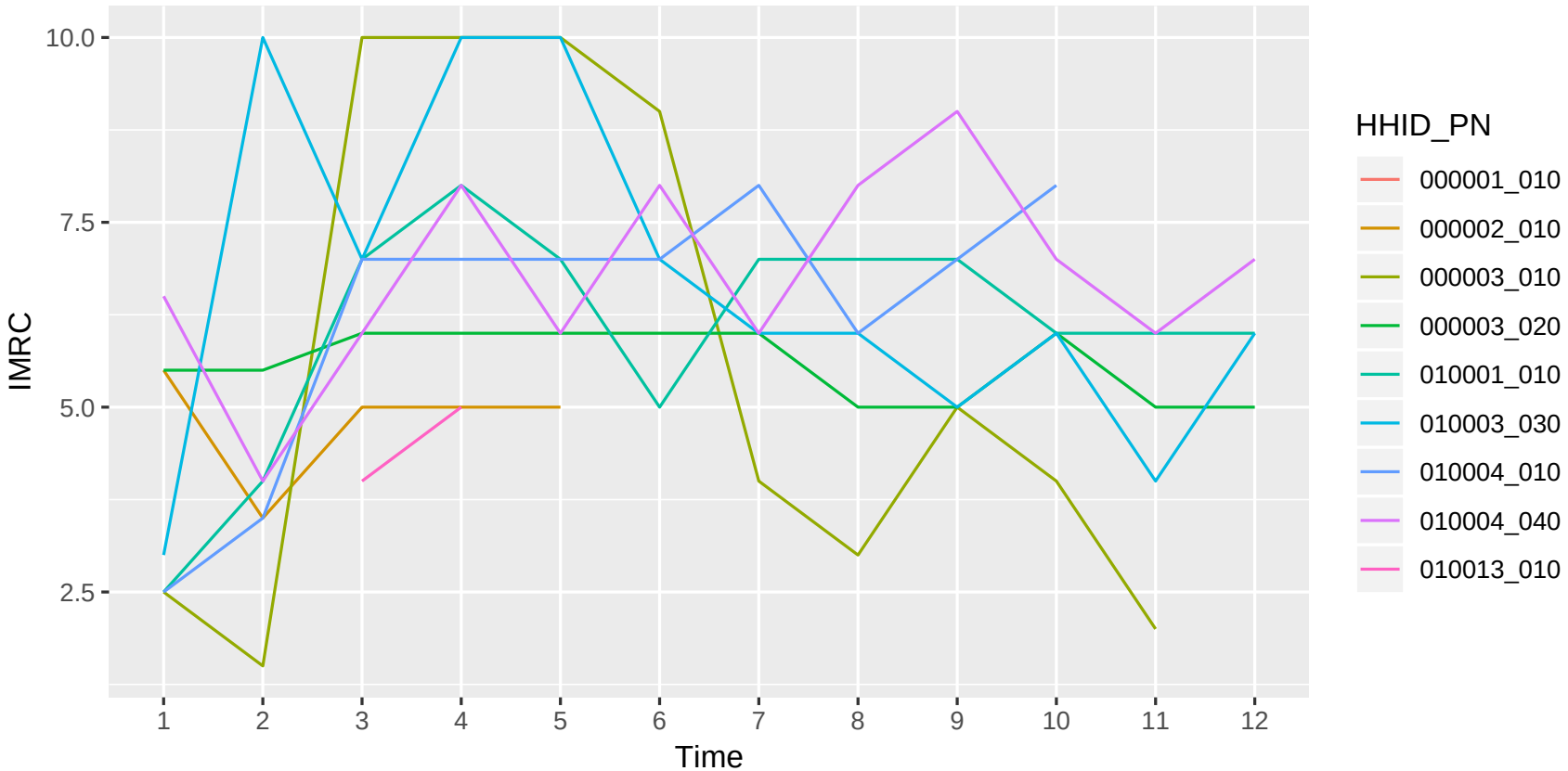
- Longitudinal study with 12 measures of approx. 20,000 probands in the USA.
- Many different variables, our focus is on cognitive performance ("Immediate Word Recall").
- Codebook: `http://hrsonline.isr.umich.edu/modules/meta/xyear/cogimp/codebook/cogimp9214a_ri.htm`

Main research question

- Development of cognitive performance over time?

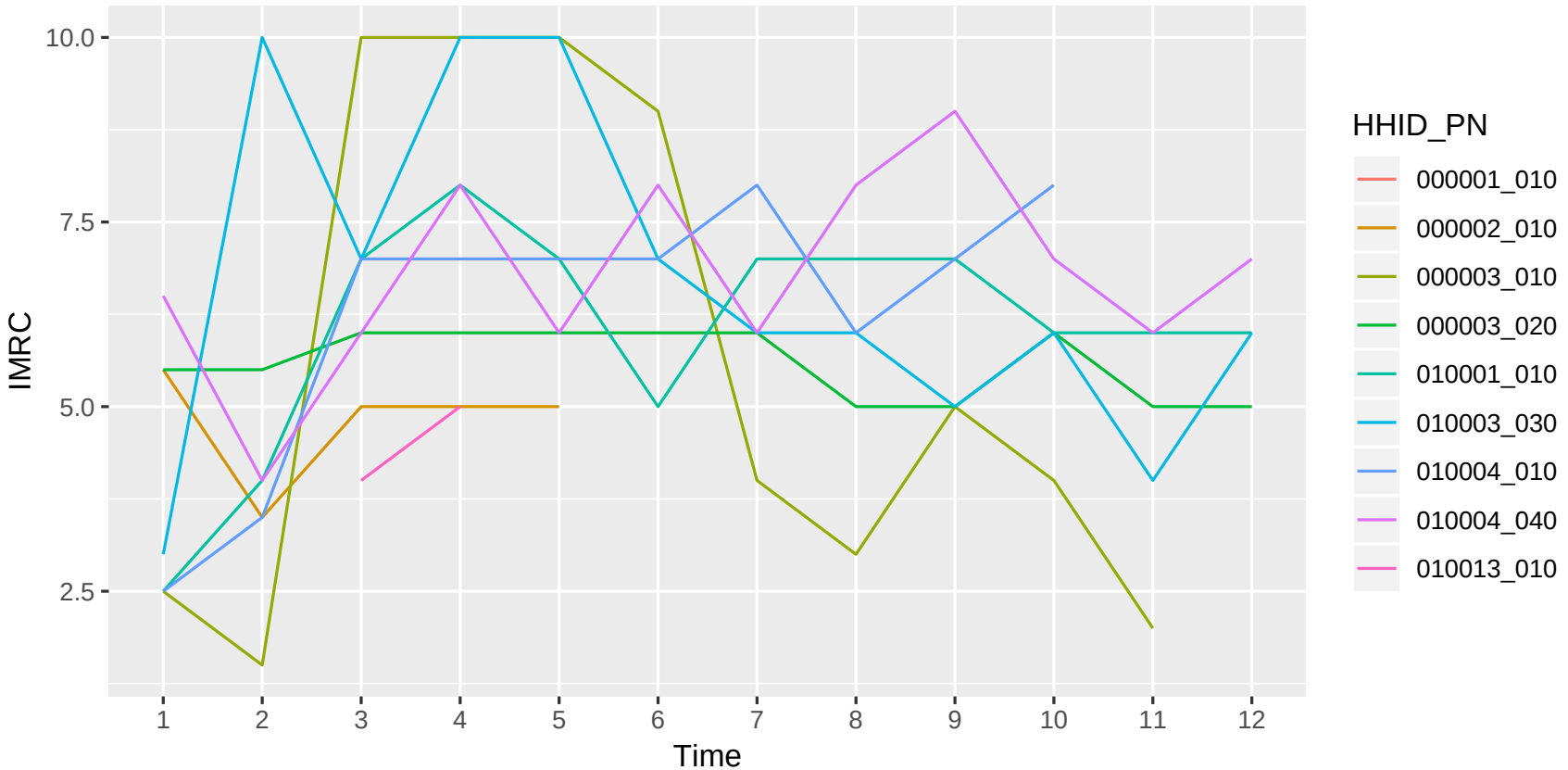
HRS (Health and Retirement Study) dataset

Trajectories of the first 10 persons in the HRS dataset (see `R_12b_HRS_data_intro.Rmd`):



HRS (Health and Retirement Study) dataset

Trajectories of the first 10 persons in the HRS dataset (see `R_12b_HRS_data_intro.Rmd`):



Creation of the figure is based on the long data format.

Insert: long and wide data format

- Long and wide formats are two different formats of datasets, where different variables of a person are either appended in columns (wide) or rows (long).
- For almost all longitudinal analyses and for some further functions (e.g. for plots) you need the long format in R.
- Transforming long to wide format can be done with several R functions, e.g. `gather` (wide $\longrightarrow$ long) and `spread` (long $\longrightarrow$ wide) in the `tidyr` package.
- See `R_12b_HRS_data_intro.Rmd`.

Insert: long and wide data format

HRS dataset in wide format:

	HHID <fctr>	PN <fctr>	HHID_PN <fctr>	T1 <dbl>	T2 <dbl>	T3 <int>	T4 <int>	T5 <int>	T6 <int>	T7 <int>	T8 <int>	T9 <int>	T10 <int>	T11 <int>	T12 <int>
1	000001	010	000001_010	5.5	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
2	000002	010	000002_010	5.5	3.5	5	5	5	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
3	000003	010	000003_010	2.5	1.5	10	10	10	9	4	3	5	4	2	*NA*
4	000003	020	000003_020	5.5	5.5	6	6	6	6	6	5	5	6	5	5
5	010001	010	010001_010	2.5	4.0	7	8	7	5	7	7	7	6	6	6
6	010003	030	010003_030	3.0	10.0	7	10	10	7	6	6	5	6	4	6

Insert: long and wide data format

HRS dataset in long format:

	HHID <fctr>	PN <fctr>	HHID_PN <fctr>	time <fctr>	IMRC <dbl>
1	000001	010	000001_010	T1	5.5
2	000001	010	000001_010	T2	*NA*
3	000001	010	000001_010	T3	*NA*
4	000001	010	000001_010	T4	*NA*
5	000001	010	000001_010	T5	*NA*
6	000001	010	000001_010	T6	*NA*
7	000001	010	000001_010	T7	*NA*
8	000001	010	000001_010	T8	*NA*
9	000001	010	000001_010	T9	*NA*
10	000001	010	000001_010	T10	*NA*
11	000001	010	000001_010	T11	*NA*
12	000001	010	000001_010	T12	*NA*
13	000002	010	000002_010	T1	5.5
14	000002	010	000002_010	T2	3.5
15	000002	010	000002_010	T3	5.0

Theoretical overview of linear mixed models

Model equation of hierarchical models

Random intercept model

$$\text{Math}_{ij} = (\beta_0 + \gamma_{0j}) + \beta_1 SES_{ij} + \beta_2 Sex_{ij} + \beta_3 Minority_{ij} + \varepsilon_{ij}$$

Random intercept, random slope model

$$\text{Math}_{ij} = (\beta_0 + \gamma_{0j}) + (\beta_1 + \gamma_{1j}) SES_{ij} + \beta_2 Sex_{ij} + \beta_3 Minority_{ij} + \varepsilon_{ij}$$

Model equation of hierarchical models

Random intercept model

$$\text{Math}_{ij} = (\beta_0 + \gamma_{0j}) + \beta_1 SES_{ij} + \beta_2 Sex_{ij} + \beta_3 Minority_{ij} + \varepsilon_{ij}$$

Random intercept, random slope model

$$\text{Math}_{ij} = (\beta_0 + \gamma_{0j}) + (\beta_1 + \gamma_{1j}) SES_{ij} + \beta_2 Sex_{ij} + \beta_3 Minority_{ij} + \varepsilon_{ij}$$

Generalization

$$Y_{ij} = (\beta_0 + \gamma_{0j}) + (\beta_1 + \gamma_{1j}) X_{1,ij} + \cdots + (\beta_k + \gamma_{kj}) X_{k,ij} + \varepsilon_{ij}$$

for observation i in cluster j, with fixed effects β and random effects γ.

Model equation of linear mixed models

Matrix notation:

$$Y = X\beta + Z\gamma + \varepsilon$$

- with fixed effects $\beta = (\beta_0, \beta_1, \ldots, \beta_k)^T$ and random effects $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_l)^T$.
- Here, the random effects γ are normally distributed and can be correlated.
- Assumption: γ and ε are uncorrelated.

Model equation of linear mixed models

Overview:

$$Y = X\beta + Z\gamma + \varepsilon$$

- with fixed effects $\beta = (\beta_0, \beta_1, \ldots, \beta_k)^T$ and random effects $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_l)^T$.
- With γ, the random effects on the cluster level can be modeled
  $\longrightarrow$ for hierarchical models.
- With ε, the dependence between the single observations (within a cluster) can be modeled
  $\longrightarrow$ for longitudinal models.

Model equation of linear mixed models

Statistical details:

$$Y = X\beta + Z\gamma + \varepsilon$$

- with fixed effects $\beta = (\beta_0, \beta_1, \ldots, \beta_k)^T$ and random effects $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_l)^T$.
- γ has a multivariate normal distribution with $l \times l$ variance-covariance matrix G: $\gamma \sim N(0, G)$. l is the number of random effects.
- ε has a multivariate normal distribution with $n \times n$ variance-covariance matrix R: $\varepsilon \sim N(0, R)$. n is the number of observations.

Model equation of linear mixed models

$$Y = X\beta + Z\gamma + \varepsilon$$

Hierarchical models

- Variance-covariance matrix R of $\varepsilon \sim N(0, R)$ is diagonal (i.e. all off-diagonal values are 0)
  $\longrightarrow$ observations within cluster are independent.

Longitudinal models

- Without cluster: variance-covariance matrix G of $\gamma \sim N(0, G)$ is diagonal $\longleftrightarrow$ no $Z\gamma$ term in the model.
- With cluster: G and R are both not diagonal (i.e. off-diagonal values are not necessarily 0).

Modeling dependencies in *R*

Overview

- Hence, R allows to model the dependencies between the different measures of a person (= observations in a cluster).
- For this, e.g. the following structures can be chosen:

Structures of R

- Unstructured: No structure of the correlation between the values of a person.
- Compound Symmetry: Correlations between all values of a person are equal (d.h. off-diagonal elements are all the same).
- Autoregressive: Correlation between the values of a person decreases over time: $r, r^2, r^3 \ldots$

Example 1

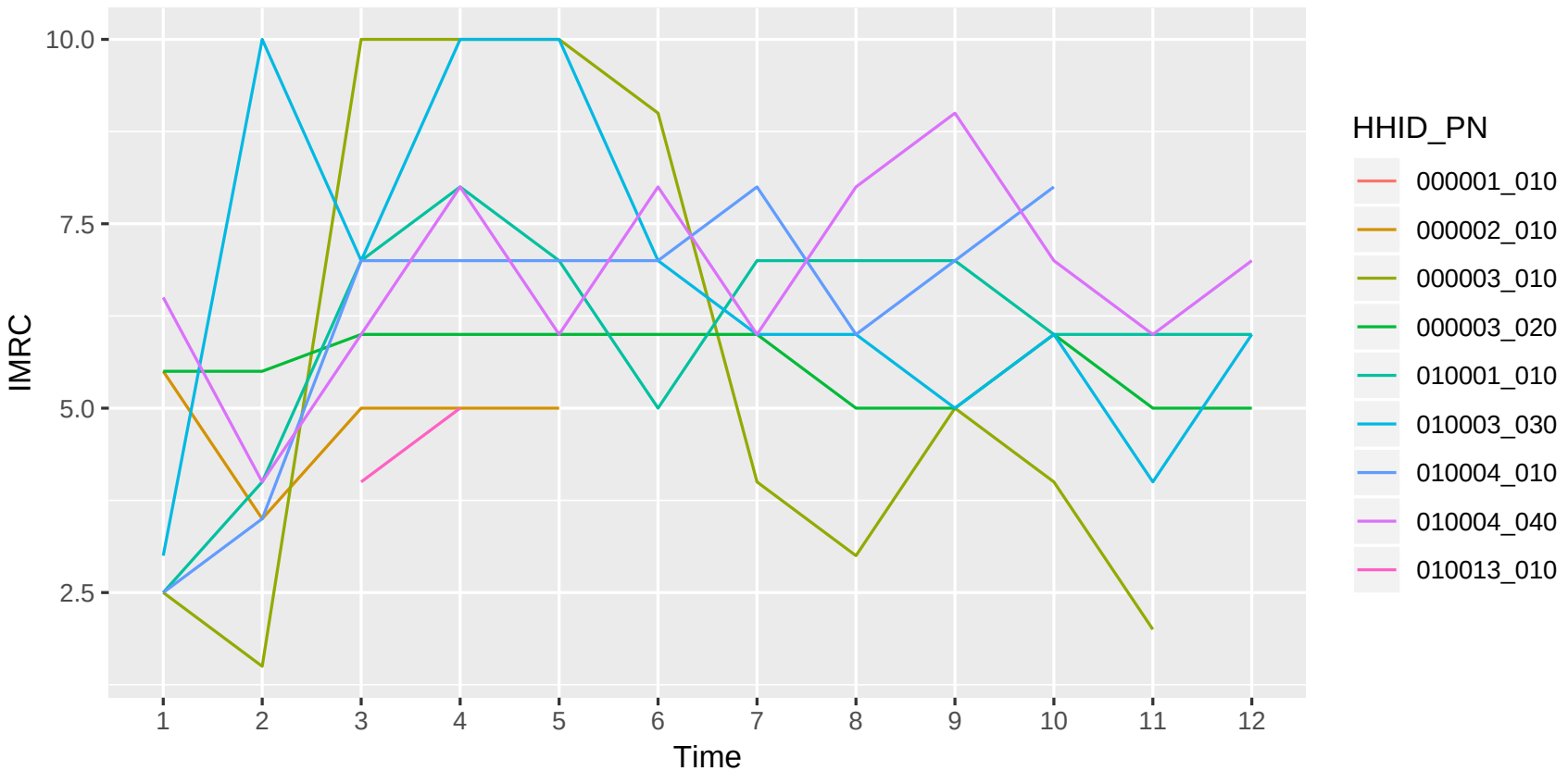
- Use the `BodyWeight` dataset in the `nlme` package in R.
- It describes the weight and diet of rats over time.
- Aim: Describe the weight trajectories over time, and predict this by diet.
- See `R_12b_example_1.Rmd`.

Example 2

- Use the `Blackmore` dataset in the `car` R package.
- It describes 138 girls with eating disorders and 98 girls from a control group, regarding their age and exercise behaviour.
- Aim: Describe the exercise behaviour over time (age) and by the group.
- See `R_12b_example_2.Rmd`.

Example 3

Aim: Describe the trajectories of cognitive performance over time in the HRS dataset.



Example 3

See `R_12b_example_3.Rmd`.

Questions?

References

- Many descriptions and tutorials online, e.g. `https://socialsciences.mcmaster.ca/jfox/Courses/soc761/Appendix-Mixed-Models.pdf`
- Agresti (2002). Categorial data analysis. Wiley.
- Bates (2010). lme4: Mixed-effects modeling with R. Springer.
- Galecki (2013). Linear mixed-effects models using R. Springer.

- `www.ccace.ed.ac.uk/research/software-resources/systematic-reviews-and-meta-analyses`.
- Handbuch für Cochrane Handbook for Systematic Reviews of Interventions: `https://training.cochrane.org/handbook`
- Pigott (2012). Advances in Meta-Analysis. Springer.
- Chen & Peace (2013). Applied Meta-Analysis with R. CRC Press.

Homework

Homework

See file `R_12_homework.Rmd`.