Biostatistics & Epidemiological Data Analysis using R

11

Linear mixed models

Stefan Konigorski

Health Intervention Analytics Group, HPI

January 18, 2023

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Content

Block	Class	Content	Date
R, Data manipulation, Descriptives	1	Overview & Introduction to R and data analysis	2022.10.19
	2	First steps in data analysis using R	2022.10.26
	3	Second steps in data analysis using R	2022.11.02
Epidemiology & Statistics: concepts	4	Epidemiological study designs	2022.11.09
	5	Estimation	2022.11.16
	6	Hypothesis testing and study planning	2022.11.23
	7	Missing data	2022.11.30
Data analysis w/ regression models	8	Linear regression I	2022.12.07
	9	Linear regression II	2022.12.14
	10	Regression models for binary and count data	2023.01.11
	11	Analysis of variance & Linear mixed models I	2023.01.18
	12	Linear mixed models II & Meta analysis	2023.01.25
	13	Survival analysis	2023.02.01
	14	Causal inference & Data analysis challenge	2023.02.08

(see full schedule online)

Mixed models - Introduction

What is a linear mixed model?

Overview

- Linear mixed model = linear model (in GLM) with fixed and random effects.
- Generalized linear mixed model (GLMM) = GLM with fixed and random effects.

Notation

- Mixed models
- Models with fixed and random effects
- Hierarchical models, multilevel analysis
- "Conditional model" (in contrast to "marginal model")
- Variance component models

What are fixed and random effects?

What are fixed effects (of predictors)?

- All models we have discussed so far had only fixed effects.
- 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.

Application of mixed models

When are mixed models relevant for data analysis?

- When observations are not independent eg when they are in clusters/hierarchical layers. eg students in schools, multiple measures of a person $\longrightarrow$ hierarchical, clustered, longitudinal observations.
- When the analysis should be adjusted for the effect of categorical covariates that have many and random categories which are not of interest. eg it is not of interest if a student in school A or school B is better, but the analysis shoul be adjusted for differences between 50 schools.
- When you are interested in interactions between variables on the individual and on the group level.
- If you are interested in a variance decomposition on multiple layers.
- If you are interested in modeling trajectories (e.g. blood pressure over time).

Application of mixed models

What can you do with mixed models?

Model dependencies and structures (clusters, hierarchical layers) in the observations, to

- consider this in the estimation of fixed effects (regression coefficients) and their standard errors + confidence intervals,
- consider this in hypothesis tests of fixed effects,
- not have to include 50 dummy variables for "school" in the model,
- investigate the variance decomposition or trajectories of Y in detail.

Application of mixed models

Which questions/data can be investigated using mixed models?

- Question: Is there an association between the age of the lecturer and the learning performance of the students?
  Data: 10 lecturers with each 15 evaluations in the class.
- Question: Is there an association between humidity and the number of accidents?
  Data: from 50 cities in 50 countries (single measurement).
- Question: Is there an association between humidity and the number of accidents?
  Data: from 50 cities in 50 countries, each 30 measurements over 1 year.

Distinguish:
Observations in clusters are independent (hierarchical).
Observations in clusters are not independent (longitudinal).

Mixed models in R

Overview

- Some mixed models can be computed using the `aov` function and `Error` option (not in the focus here).
- The two most used R functions for mixed models are `lme` in the `nlme` (nonlinear mixed effects) package and `lmer` in the `lme4` (linear mixed effects with S4 classes) package.

Comparison

- Many models can be computed with both functions yielding the same results.
- `lmer`: newer, faster, no p-values for fixed effects.
- `lme`: older, better documentation, p-values for fixed effects, can model more correlation and variance structures.

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?

$\longrightarrow$ In the following, we will look at different models and their interpretation in the `MathAchieve` dataset.

$\longrightarrow$ see `R_11b_exercise_LMM.Rmd`

Intercept-only model

Model

- Predict the grade of each child $i = 1 \dots n$ by the general mean:
$$\text{Math}_i = \beta_0 + \varepsilon_i$$
- This is a normal linear regression without predictors (only β_0 as fixed effect).

Which assumptions does this model contain?

- The math grade is the same in all students.

$\longrightarrow$ Is this assumption realistic (i.e. fits the data)?

in R

- `lm(MathAch ~ 1, data = MathAchieve)`

Fixed-effect (only) model

Model

- Predict the grade of each child $i = 1 \ldots n$ by the general mean and SES:
$$\text{Math}_i = \beta_0 + \beta_1 SES_i + \varepsilon_i$$
- This is still a regular linear regression (with β_0, β_1 as fixed effects).

Which assumptions does this model contain?

- The math grade only depends on the SES and is the same in all students with the same SES.

$\longrightarrow$ Is this assumption realistic (i.e. fits the data)?

in R

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

Fixed-effects (only) model

Model

- Predict the grade of each child $i = 1 \dots n$ 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 linear 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.

$\longrightarrow$ Is this assumption realistic (i.e. fits the data)?

in R

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

Interim evaluation

Questions

- Can the models so far predict the math grade well?
- Does school play a role?
- If yes, how can we incorporate this in the analysis?

How can school be considered in the analysis?

Possibility 1

- Include dummy variable for every school in regression/ANOVA

Assumptions

- How is school considered? Which assumption is made on the association between SES and math grade?
- Assumption: the association between SES and math grade is the same in all schools and only differs by a constant (i.e. same slope of the regression line, different intercept).

Model

$$\text{Math}_i = \beta_0 + \beta_1 SES_i + \beta_2 Sex_i + \beta_3 Minority_i + \sum_j \beta_{4j} School_{ij} + \varepsilon_i \text{ or}$$

$$\text{Math}_{ij} = \beta_0 + \beta_{4j} + \beta_1 SES_{ij} + \beta_2 Sex_{ij} + \beta_3 Minority_{ij} + \varepsilon_{ij} \quad \text{in school j.}$$

How can school be considered in the analysis?

Possibility 1

- Include dummy variable for every school in regression/ANOVA

Assumptions

- How is school considered? Which assumption is made on the association between SES and math grade?
- Assumption: the association between SES and math grade is the same in all schools and only differs by a constant (i.e. same slope of the regression line, different intercept).

Consequences/Summary

- Consequence: loss of degrees of freedom and power!
- But: if differences between schools are of interest, then this is the way to go!

How can school be considered in the analysis?

Possibility 2

- Include school as continuous variable in the regression.

Assumptions

- How is school considered? Which assumption is made on the association between SES and math grade?
- Assumption 1: Math grade increases constantly between ordered school (nonsense!).
- Assumption 2: Association between SES and math grade is same in all schools and only differs by a constant ("fixed increasing") constant (also nonsense!).

How can school be considered in the analysis?

Possibility 3

- Stratified analysis, separately for each school.

Disadvantages?

- Small sample size, low power.
- No aggregated results.
- Variables on school level cannot be considered.

How can school be considered in the analysis?

Possibility 4

- Analysis not on individual but school level.
- I.e. aggregate all variables on school level (compute means), then do regression with these observations.

Disadvantages?

- Small sample size.
- Results have to be interpreted on the school level, not on the individual level (ecological fallacy!).

$\longrightarrow$ Still these are all models with fixed effects!

How can school be considered in the analysis?

Possibility 5

- 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.

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

$\longrightarrow$ look at how much variance can be explained: variance components.

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_11b_exercise_LMM.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.