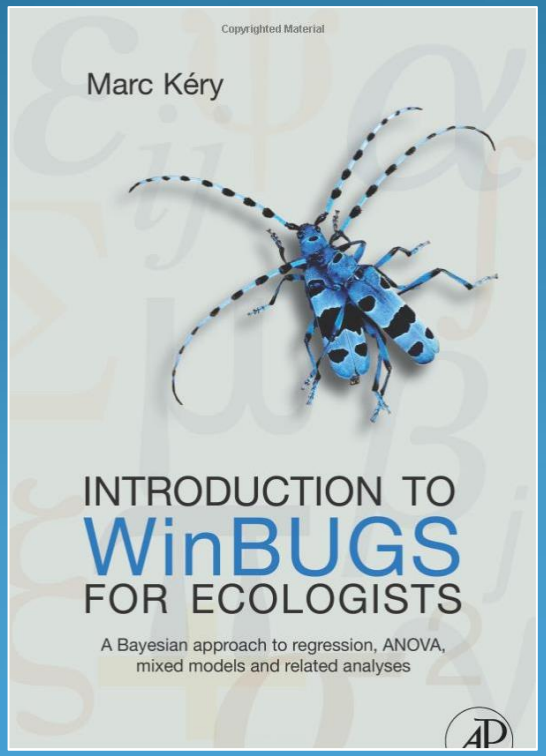
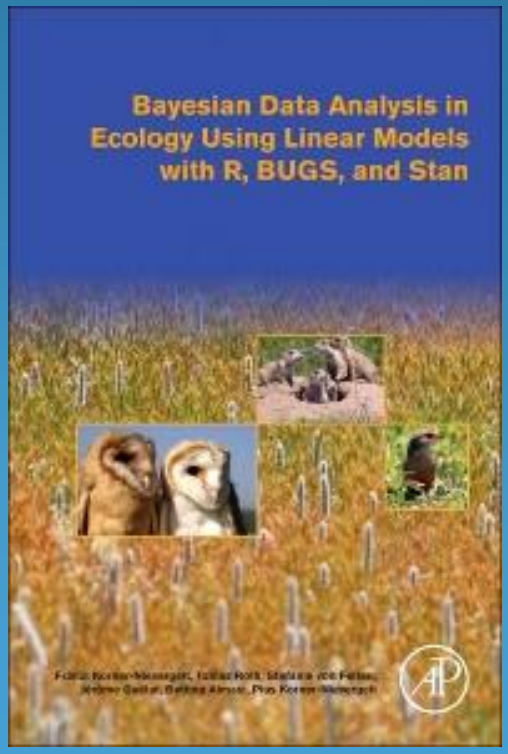
Topic 7: Linear mixed-effects models (random effects)



Chapter 12



Chapter 7

Recap: (simple) normal linear regression

lm(mass ~ svl)

$$\mathbf{mass}_i = \boldsymbol{\alpha} + \boldsymbol{\beta} * \mathbf{svl}_i + \boldsymbol{\varepsilon}_i$$

α = constant

β = constant to be multiplied with snout–vent leng.

ε_i = residual for snake i

$\varepsilon_i \sim$ Normal(0,σ^2)

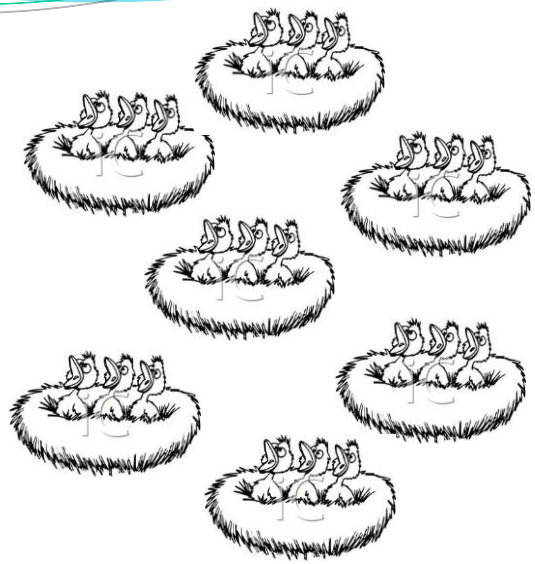
„The core of modern applied statistics“

	Single Random Process	Two or More Random Processes
Normal response	Linear model (LM)	Linear mixed model (LMM)
Exponential family response	Generalized linear model (GLM)	Generalized linear mixed model (GLMM)

Linear mixed-effects models

- Mixed models = hierarchical models
- Contain factors with both fixed and random effects
- The presence of a random effect adds another layer of variance!

Example



- Measure growth rates of nestlings in different nests
- Mass measurements of each nestling several times during the nestling phase

- Measurements from the same individual are likely to be more similar (grouped within nestlings - repeated measurements)
- Individuals from the same nest are likely to be more similar (grouped within nests)

If the grouping structure of the data is ignored in the model, the residuals do not fulfil the independence assumption!!

Fixed or random effects?

- **Fixed effects** (as factors) have a finite ("fixed") often low number of levels e.g. "sex"
- **Random effects** have a theoretically infinite number of levels of which we have measured a random sample. E.g. we can measure 5 or 10 or n nests
- For **fixed effects** we are interested in the specific differences between levels (e.g., between males and females)
- For **random effects** we are only interested in the between-level (= between-group, e.g., between-nest) variance rather than in differences between specific levels (e.g., nest A versus nest B)

It depends sometimes on the aim of the study whether a factor should be treated as fixed or random.

Fixed or random effects?

Fixed factors

- Groups are predetermined, of direct interest, repeatable
- e.g. treatment, sex, age class, season, habitat

Conclusions about differences among groups can only be applied to the groups of the study and cannot be generalized to other treatments, habitats…

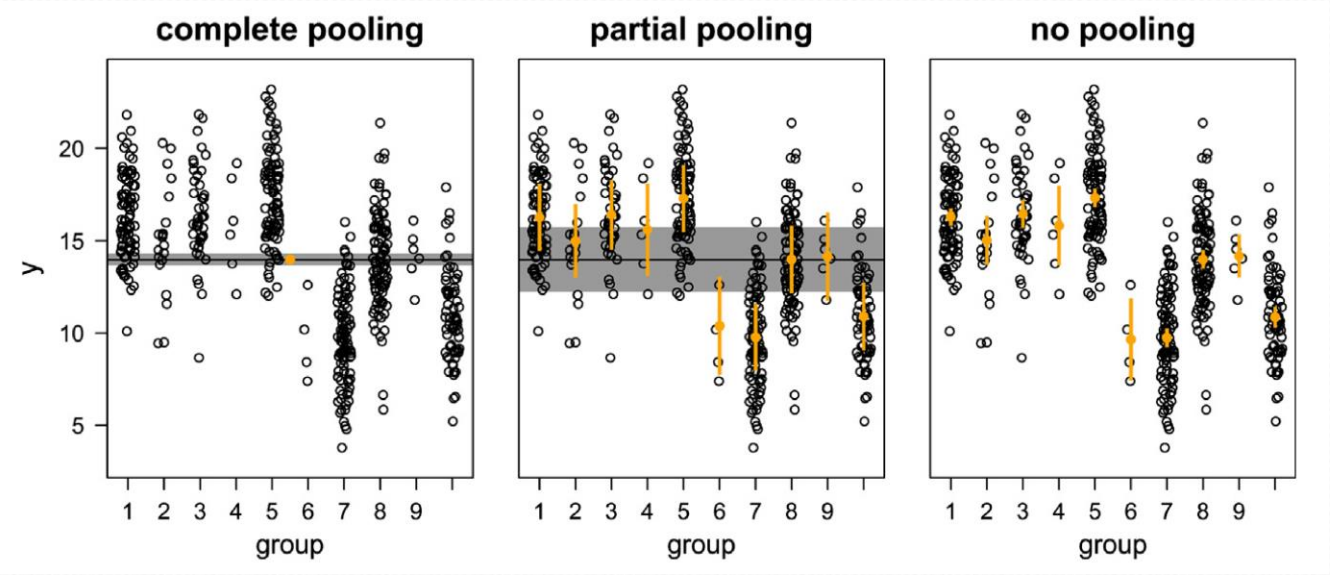
Random factors

- Groups are a random sample from a population of groups
- e.g. individual, nest, field, school, study plot

Conclusions about differences among groups **can** be generalized to the whole population.

The variance between groups is the main interest not the specific group means!

3 ways to obtain means of grouped data

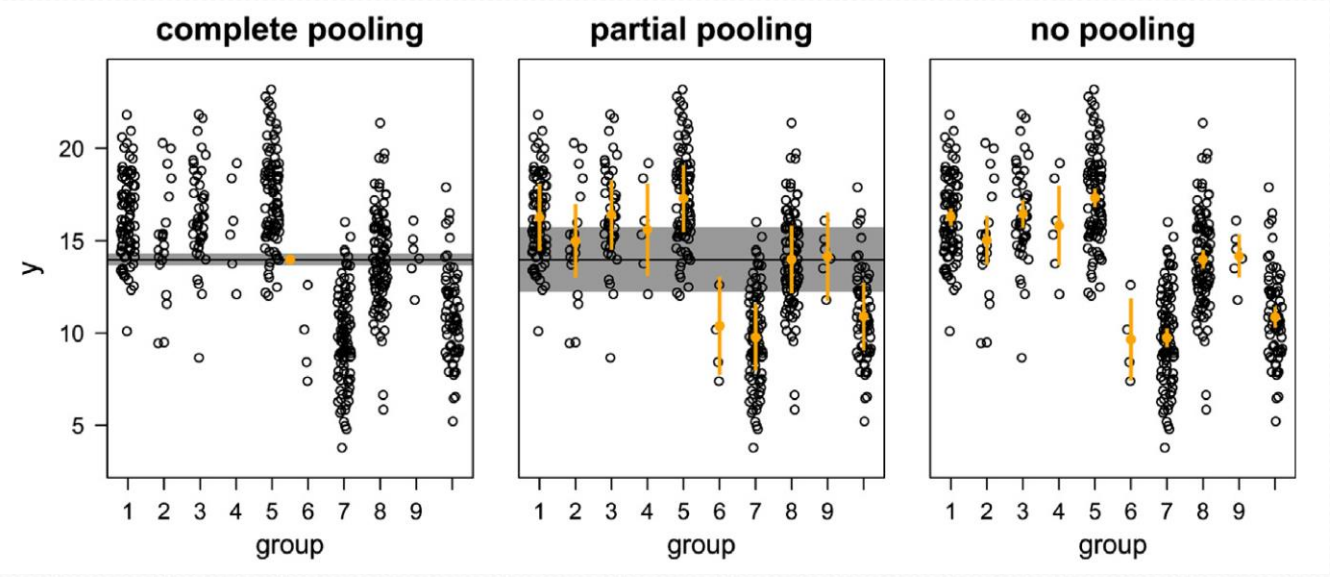


1. Grouping is ignored – **complete pooling**

-> Pseudpreplication!

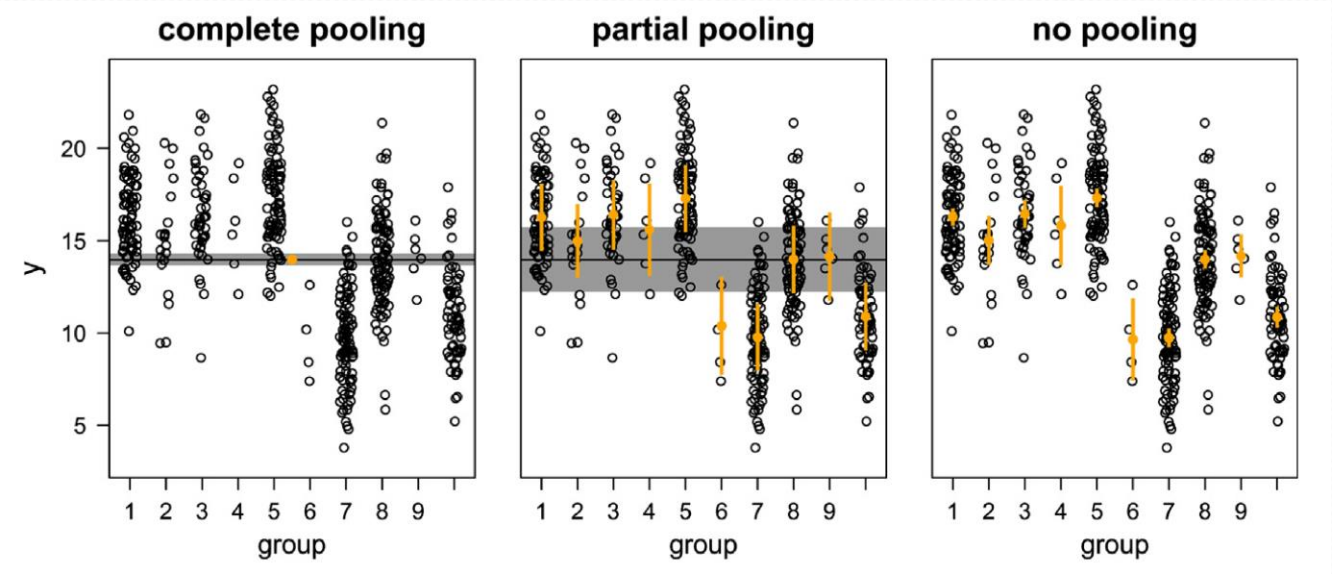
we are too confident in the result because we assume that all observations are independent when they are not.

3 ways to obtain means of grouped data



2. Group means are estimated separately - **no pooling**
(data from all other groups are ignored when estimating a group mean - equivalent to treating the factor as fixed)
-> danger of overestimation of the *between-group variance* because the group means are estimated independently of each other

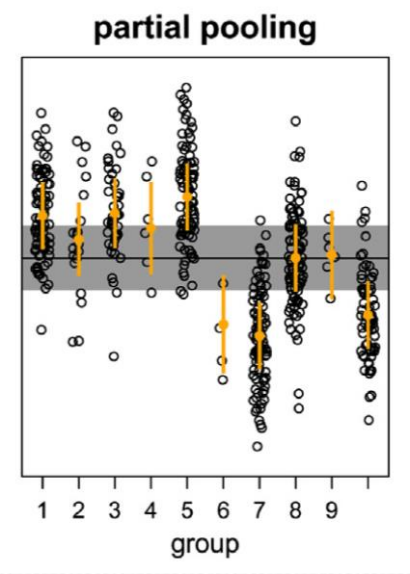
3 ways to obtain means of grouped data



3. Partially pooling - group means are weighted averages of the population mean and the unpooled group means. The weights are proportional to sample size and the inverse of the variance.

Advantages of partial pooling

- Assumes that the group means are a random sample from a common distribution
- Information is exchanged between groups.
- Estimated means for groups with low sample sizes, large variances, and means far away from the population mean are shrunk toward the population mean



Linear mixed-effects models

- To summarize: mixed models are used to appropriately estimate between-group variance and to account for non-independency among data points.
- The presence of a random effect adds another layer of variance!

$$y_i = \alpha_{j(i)} + \beta_{j(i)} * x_i + \varepsilon_i$$

$$(\alpha_j, \beta_j) \sim \mathrm{MVN}(\mu, \boldsymbol{\Sigma}) \quad \text{\# Bivariate normal random effects}$$

$$\mu = (\mu_\alpha, \mu_\beta) \quad \text{\# Mean vector}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_\alpha^2 & \sigma_{\alpha\beta} \\ \sigma_{\alpha\beta} & \sigma_\beta^2 \end{pmatrix} \quad \text{\# Variance–covariance matrix}$$

$$\varepsilon_i \sim \mathrm{Normal}(0, \sigma^2) \quad \text{\# Residual “random” effects}$$

Fitting linear mixed-effects models in R

-> use package lme4, function: **lmer()**
(lm() assumes all effects are fixed)

- lmer() is used similarly to function lm()
- random factors are added in the model formula within parentheses.
- The "1" stands for the intercept and the "|" means "grouped by", (1|Individual) therefore, adds the random deviations for each individual to the average intercept.

Interpretation of the R output

Example: species richness on five plots at each of nine beaches with the predictor variable NAP (height of sampling station compared to mean tidal level). Data collected by the dutch institute RIKZ, see also Zuur et al. 2009

Interpretation

- Fitted with REML
- Parameter estimates are grouped into a random effect and fixed effects.

```
Linear mixed model fit by REML ['lmerMod']
Formula: Richness ~ NAP + (1 | Beach)
   Data: dat

REML criterion at convergence: 239.5

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.4227 -0.4848 -0.1576  0.2519  3.9794

Random effects:
 Groups   Name        Variance Std.Dev.
 Beach    (Intercept) 8.668    2.944
 Residual             9.362    3.060
Number of obs: 45, groups:  Beach, 9

Fixed effects:
            Estimate Std. Error t value
(Intercept)   6.5819     1.0958   6.007
NAP          -2.5684     0.4947  -5.192

Correlation of Fixed Effects:
    (Intr)
NAP -0.157
```

Interpretation of the R output

- Random effects section gives the estimates for the between beach variance = 8.7, and the residual variance = 9.4
- Fixed effects section gives the estimates for the population ("average beach") intercept =6.6 and the slope parameter for NAP = -2.6
- No R^2-> see r.squaredGLMM in MuMIn package marginal and conditional R^2 Nakagawa & Schielzeth 2013

```
Linear mixed model fit by REML ['lmerMod']
Formula: Richness ~ NAP + (1 | Beach)
   Data: dat

REML criterion at convergence: 239.5

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.4227 -0.4848 -0.1576  0.2519  3.9794

Random effects:
 Groups   Name        Variance Std.Dev.
 Beach    (Intercept) 8.668    2.944
 Residual             9.362    3.060
Number of obs: 45, groups:  Beach, 9

Fixed effects:
            Estimate Std. Error t value
(Intercept)   6.5819     1.0958   6.007
NAP          -2.5684     0.4947  -5.192

Correlation of Fixed Effects:
    (Intr)
NAP -0.157
```

Interpretation

- No p-values!
- The estimation of standard errors for the model parameters in mixed models is difficult using frequentist methods
- The problem is that in mixed models it is difficult to obtain the degrees of freedom.
- However, for testing fixed effects, and when sample size is large, the approximate likelihood ratio test is reliable in practice.
- In contrast, when testing random effects, or when sample size is small, the approximate likelihood ratio test can be misleading.

- Solutions: lmerTest („quick and dirty“ with likelihood ratio test), bootstrapping, MCMC (bayesian method)

REML

- Default for a mixed model: the restricted maximum likelihood estimation (REML) instead of the maximum likelihood (ML)
- ML method underestimates the variance parameters because it assumes that the fixed parameters are known without *uncertainty* when estimating the variance parameters.
- When sample size is large compared to the number of model parameters, the differences between the ML and REML estimates become negligible.
- As a guideline, use REML if the interest is in the random effects (variance parameters) and ML if the interested is in the fixed effects.
- The estimation method can be chosen by setting the argument “REML” to “FALSE” (default is “TRUE”).

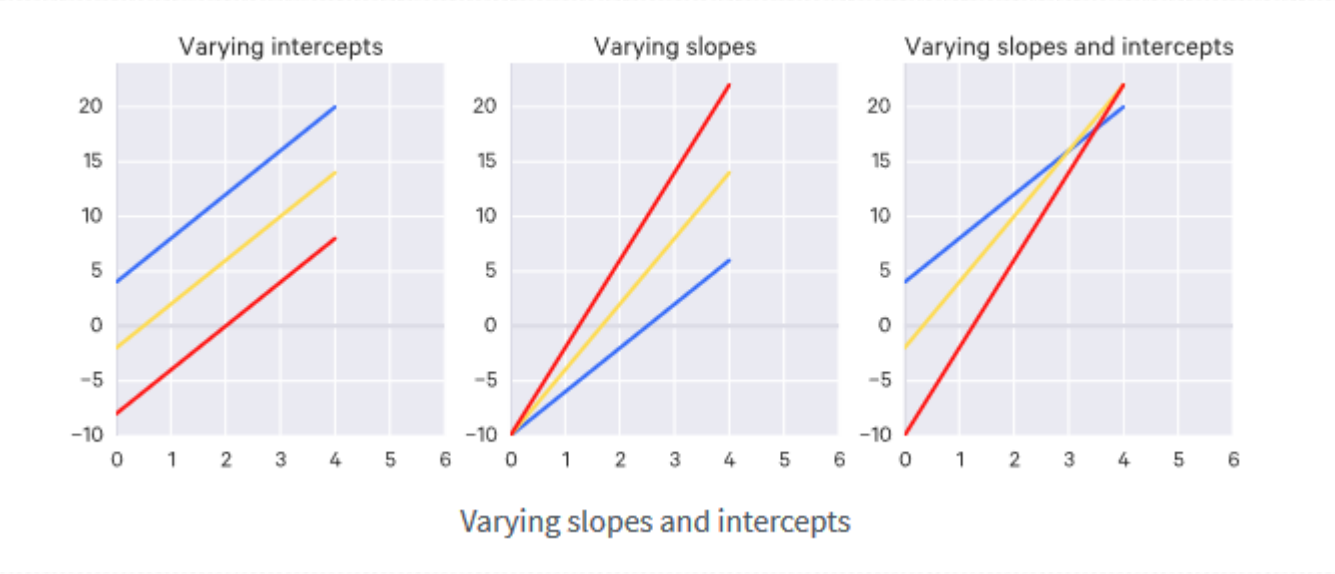
Assumptions

- In principle, the same methods described for linear models are used to assess violation of model assumptions in mixed models.

- plot(model) does not work here - functions have to be coded by hand

- Plot the residals versus fitted values and the QQ plot for both the fixed and the random effects!
  Both error terms have to be normally distributed!

Random intercept and/or slope



- So far only the intercept α was modeled per individual (the model allowed for between-individual variance)
- The random effect does not need to be restricted to the intercept!
- We cannot include an individual-specific difference e.g. individuals might react differently to a treatment