Linear Mixed Models (LMM)

Emily Haeuser

Eva Malecore

Ecology Konstanz

TReND

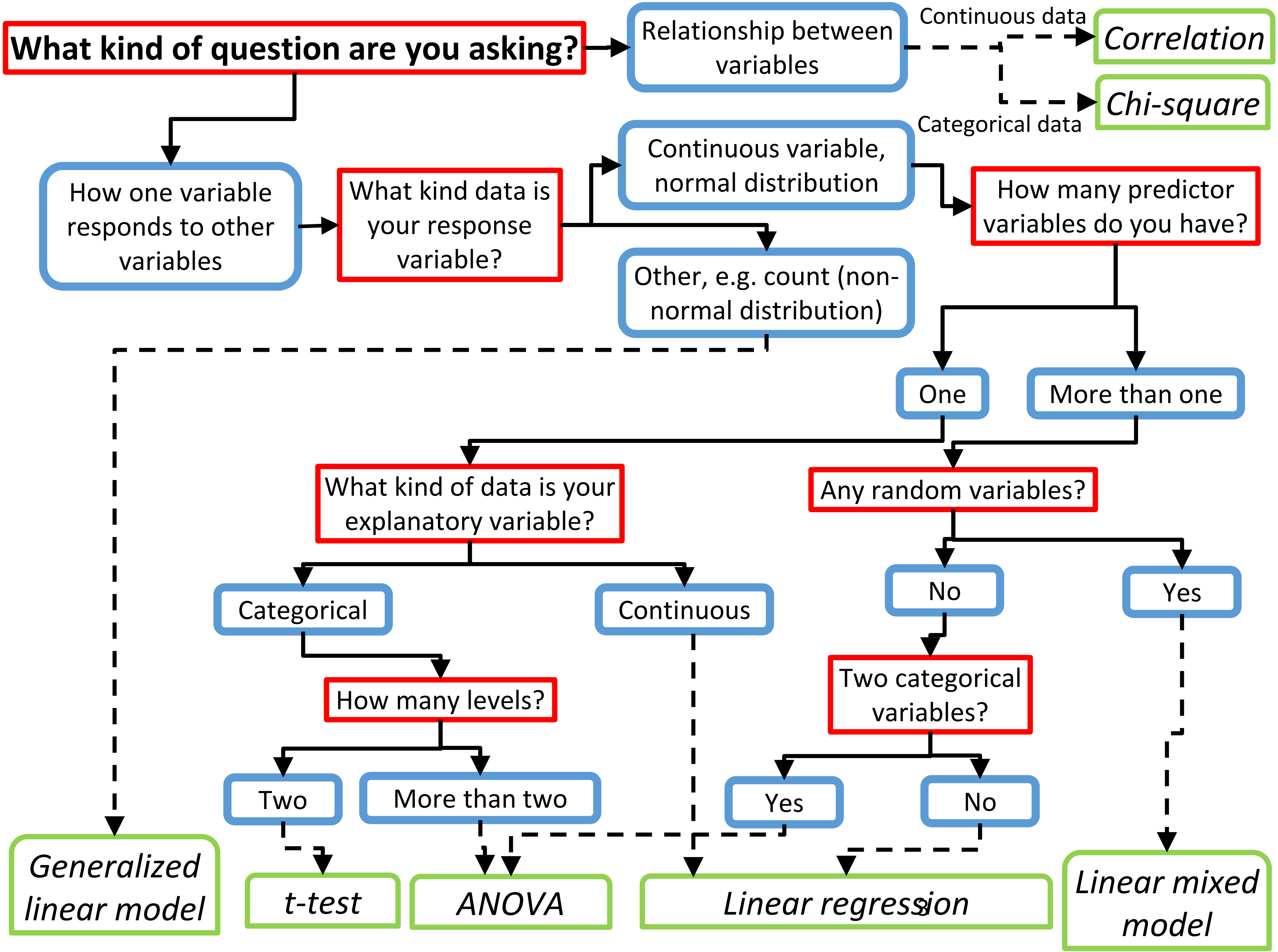
Institute for Health Metrics and Evaluation

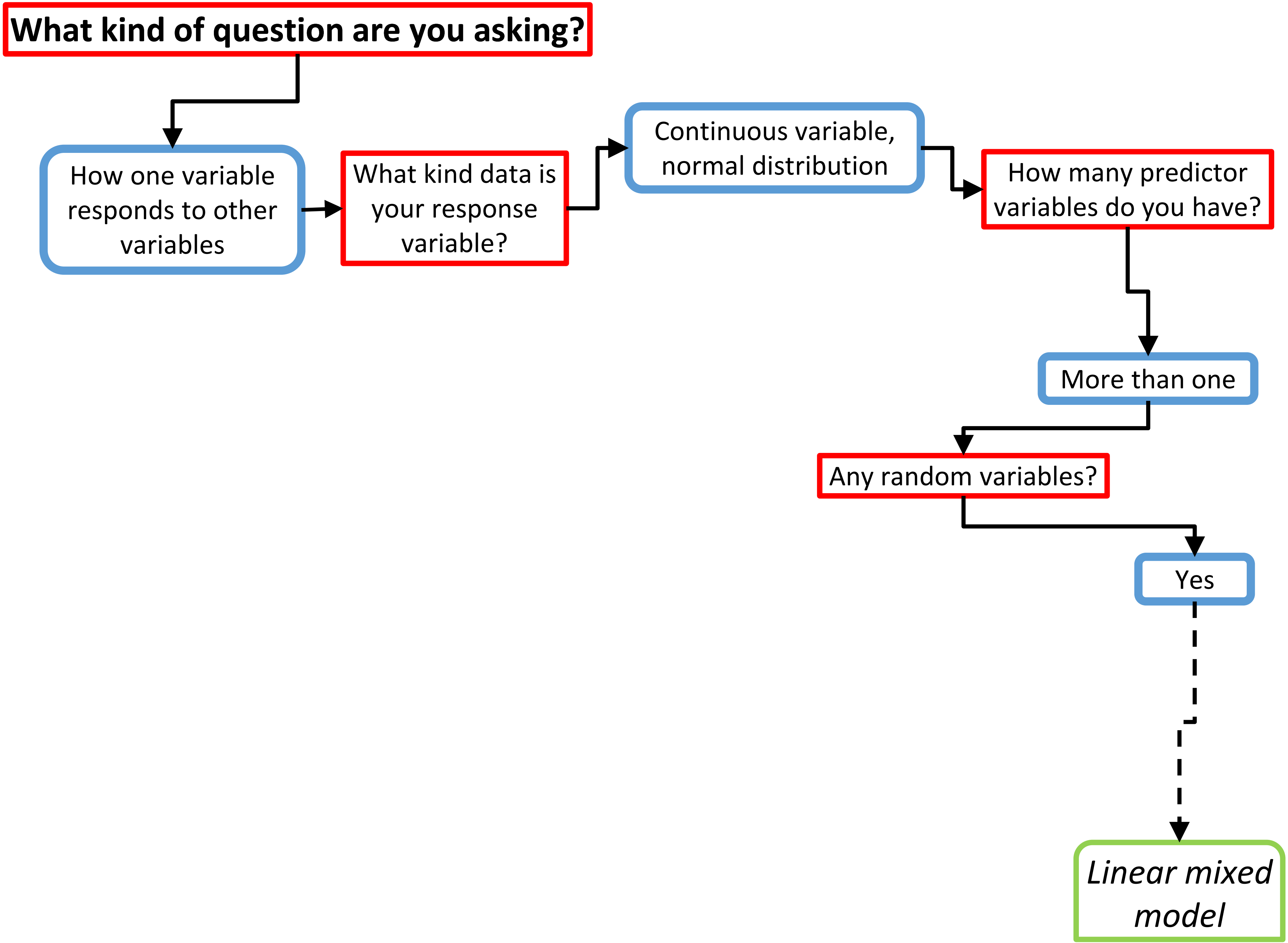
The Company of Biologists

International Max Planck Research School for Organismal Biology

Overview

Topics
1 – Random Effects
2 – How to fit linear mixed-effect models (LMM)
3 – Random slope, corssed and nested random effects


What kind of question are you asking?
Relationship between variables
Continuous data
Correlation
Chi-square
Categorical data
How one variable responds to other variables
What kind data is your response variable?
Continuous variable, normal distribution
Other, e.g. count (non-normal distribution)
How many predictor variables do you have?
One
More than one
What kind of data is your explanatory variable?
Any random variables?
Categorical
Continuous
No
Yes
How many levels?
Two categorical variables?
Two
More than two
Yes
No
Generalized linear model
t-test
ANOVA
Linear regression
Linear mixed model



What kind of question are you asking?

How one variable responds to other variables

What kind data is your response variable?

Continuous variable, normal distribution

How many predictor variables do you have?

More than one

Any random variables?

Yes

Linear mixed model

Random Effects

In all previous examples, we have treated all **categorical explanatory variables** as if they were the same

There are actually two different sorts of categorical explanatory variables: **fixed effects** and **random effects**.

So-called '**mixed models**' contain both fixed and random effects.

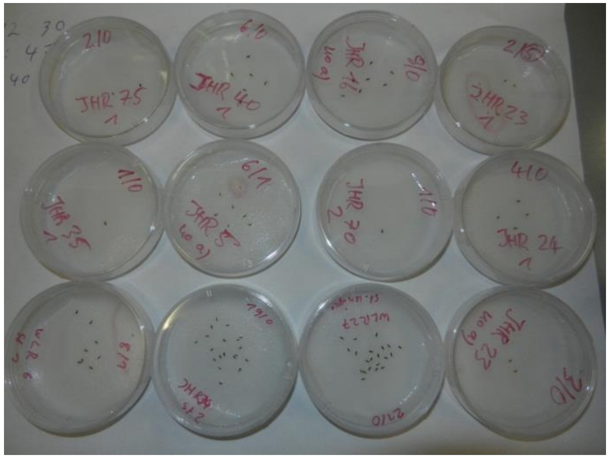
A **factor** is fixed when the levels under study are the only levels of interest.

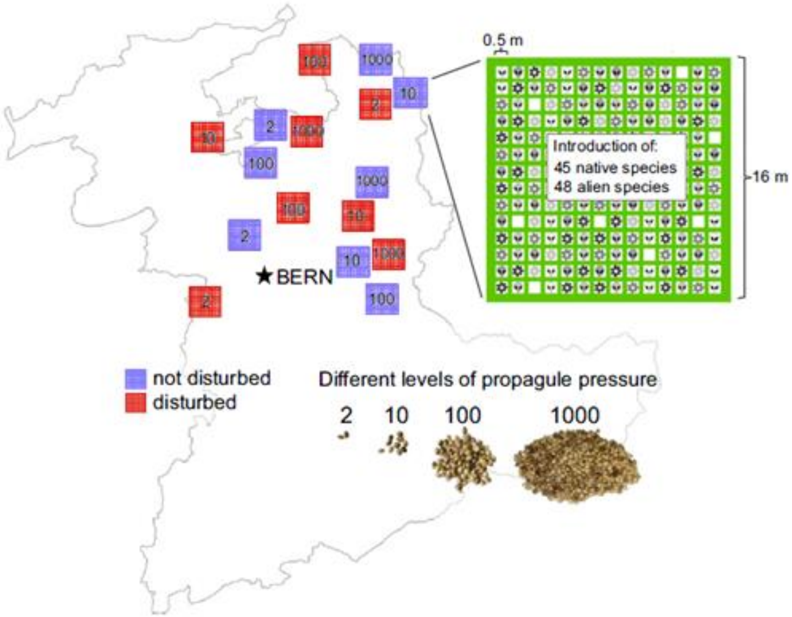
A **factor** is **random** when the levels under study are a **random** sample from a larger population and the goal of the study is to make a statement regarding the larger population.

Random Effects

Non independent or grouped data, hierarchical data:

- Lack of independence between data points
  - Repeated measurements on the same individual
- Data may be grouped into experimental blocks
  - Blocks may coincide with some extra, unmeasured variable
  - Some variation in experiment may be explained by block effects
- Data may have a hierarchical structure
  - -Block within plot





Fixed vs Random Effects

Fixed effects should:

- be variables for which **we are interested in differences between levels**

e.g. effects of fertilization level

- have **levels that are non-random**

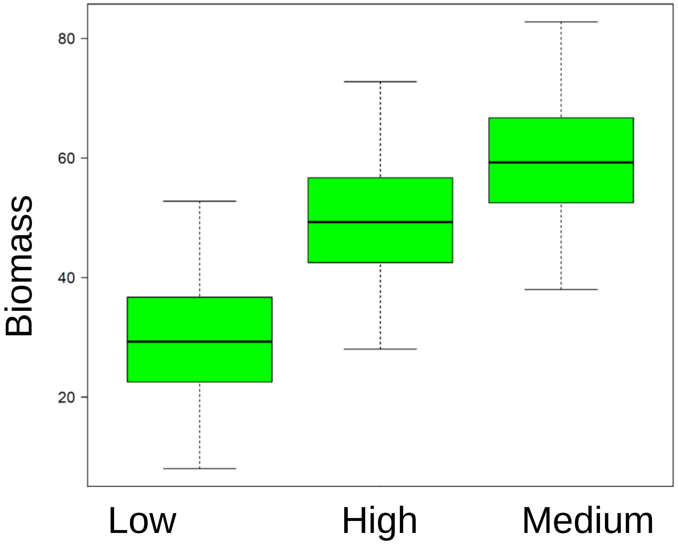
e.g. levels of fertilization were specifically chosen and set

- have **few levels**

e.g. there are only 3 fertilization levels

- have **levels which are informative**

e.g. medium fertilization means more fertilizer than low, and less than high



Fixed vs Random Effects

- **Random effects** should:
  - be variables for which **we are not interested in effect sizes, but only in variation among levels**

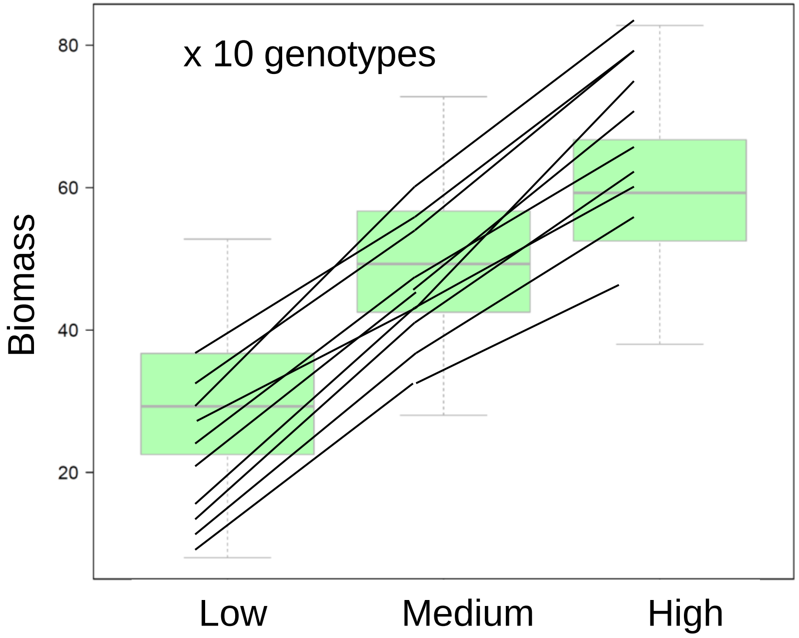
e.g. if we had 10 genotypes, in order to generalize and account for genotypic variation

  - have **levels that are randomly chosen from a 'population' of levels**

e.g. there was no a priori *reason why these particular genotypes were chosen (= no bias)*

  - have **many levels** *(a threshold for 'many' is subjective…)*
  - have **levels which are uninformative**

e.g. if there is no bias, one genotype should be as different from the others, on average, as the next one, i.e. being one genotype or another informs nothing

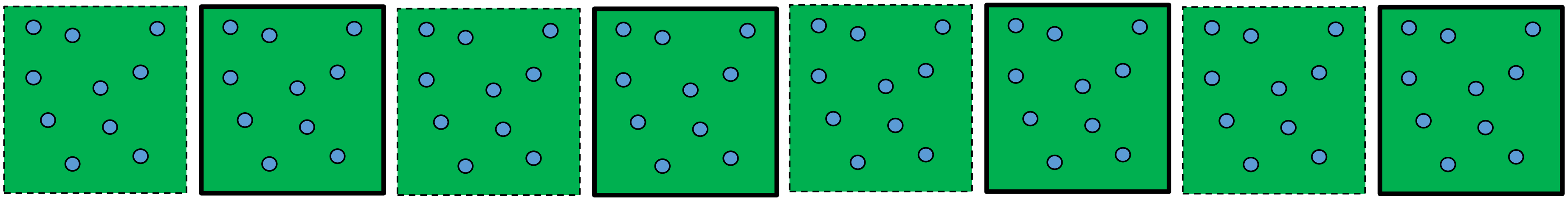


Examples

Examples (1/4)

Question: influence of herbivory on soul nutrient levels

Experiment: We exclude deer from a patch of vegetation, to assess effects of herbivory on soil nutrient levels (x 4 plots of each treatment). We take 10 soil cores from each plot.



Our data points are not independent, treating them as such is what we call pseudoreplication!

Our unit of observation is replicated within units of the treatment replicate: we have a **nested design**.

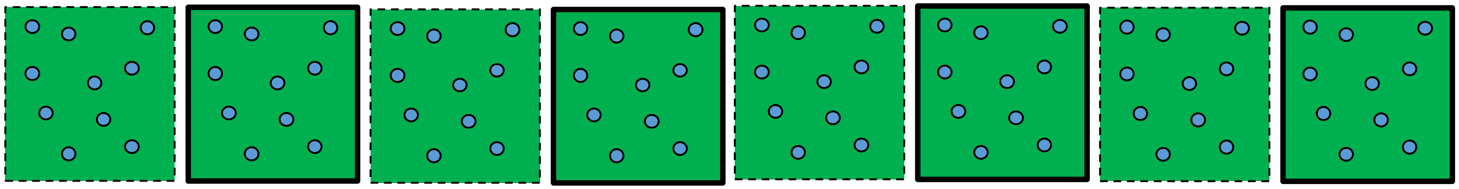
Examples

fixed effects	random effects
the levels (not the effects!) are fixed	the levels are not fixed
differences between specific levels are of interest	variance between the levels is of interest (or needs to be accounted for in the model)

Example (1/4)

- Treatment (**herbivory excluded**/ not excluded)

- Plot

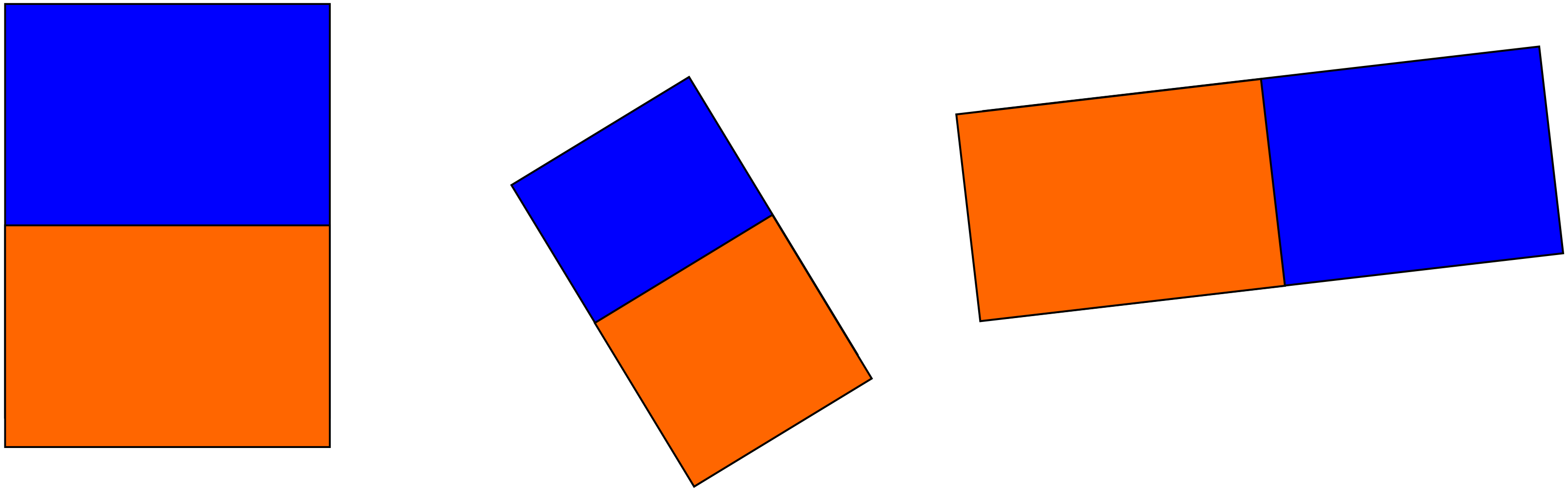


Examples

Examples (2/4)

Question: difference in biodiversity between different managements in agricultural fields

Experiment: apply management A and B in half of each field

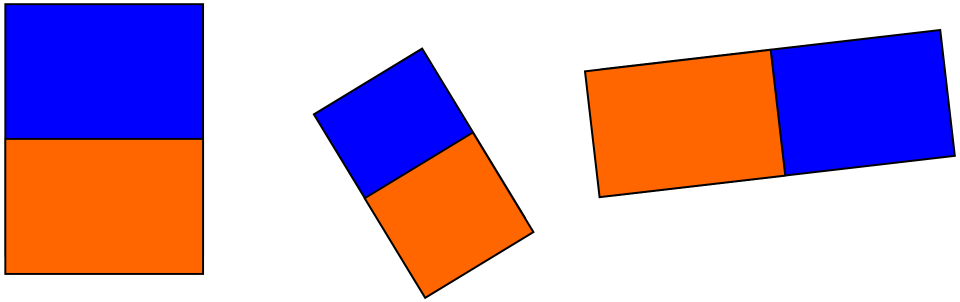


Examples

fixed effects	random effects
the levels (not the effects!) are fixed	the levels are not fixed
differences between specific levels are of interest	variance between the levels is of interest (or needs to be accounted for in the model)

Example (2/4)

- Management (A or B)
- Field

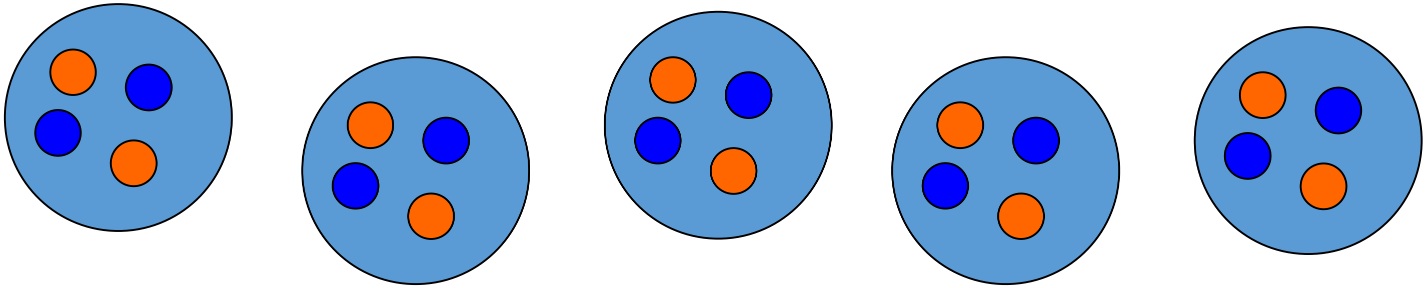


Examples

Examples (3/4)

Question: influence of corticosterone-implant on barn owl nestling growth rate, measured multiple times on same individual

Experiment: in each of n nests,
2 nestlings with corticosterone implant
2 netslings with a placebo



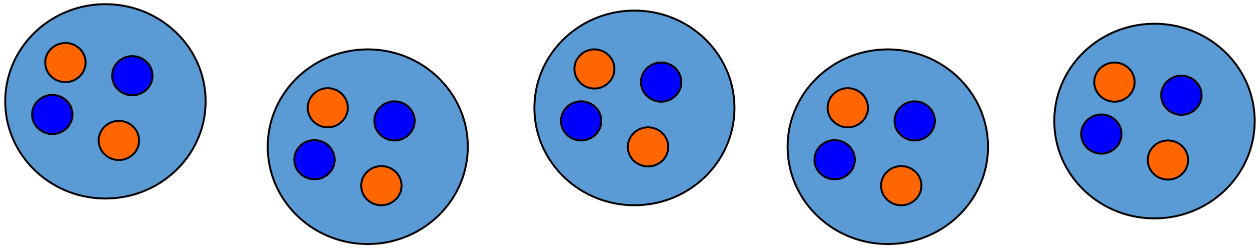
Examples

fixed effects	random effects
the levels (not the effects!) are fixed	the levels are not fixed
differences between specific levels are of interest	variance between the levels is of interest (or needs to be accounted for in the model)

Example (3/4)

- Implant (Placebo or Corticosterone)

- Nest
- Individual

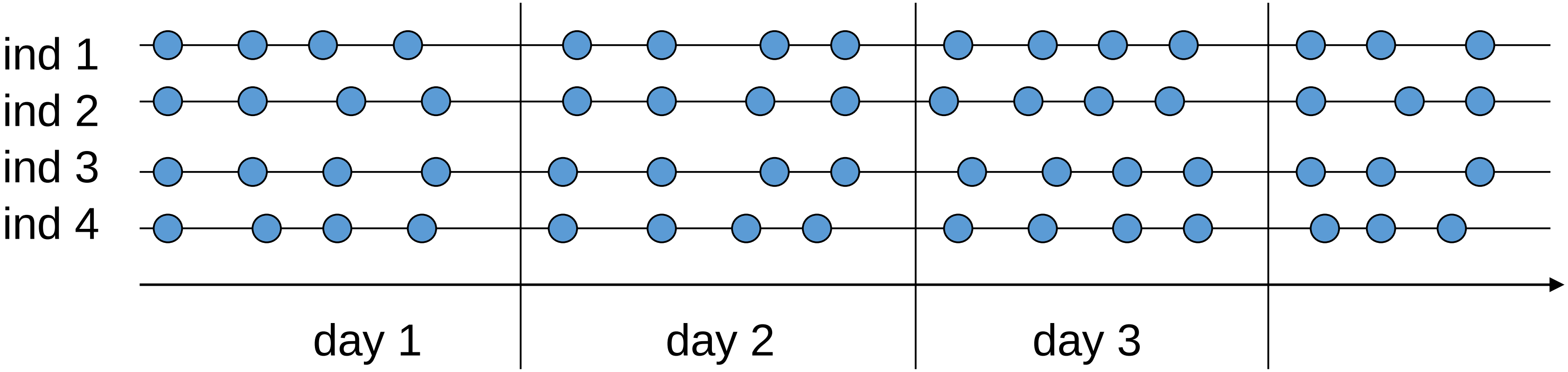


Examples

Examples (4/4)

Question: is the behaviour of an animal influenced by weather (means per day) and time of day?

Experiment: observation of n individuals four times a day



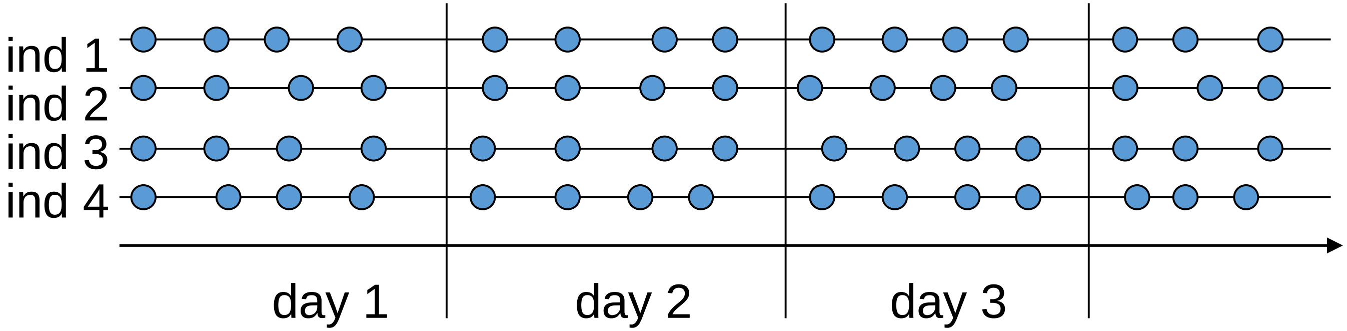
Examples

fixed effects	random effects
the levels (not the effects!) are fixed	the levels are not fixed
differences between specific levels are of interest	variance between the levels is of interest (or needs to be accounted for in the model)

Example (4/4)

- Time of the day
- Mean daily temperature

- Day
- Individual



Random Effects

- **Random effects** are a way of grouping the data that we are not interested in.
- But the key thing is that **they make our data points non-independent!**
- And remember, **independence between data points is a key assumption** of all the standard statistical tests and models we have seen so far. So, we need to account for those random effects.
- Such lack of independence between data points due to random effects can occur in both **observational and experimental studies**. Hence the **importance of the design** of the field sampling or of the experiment!

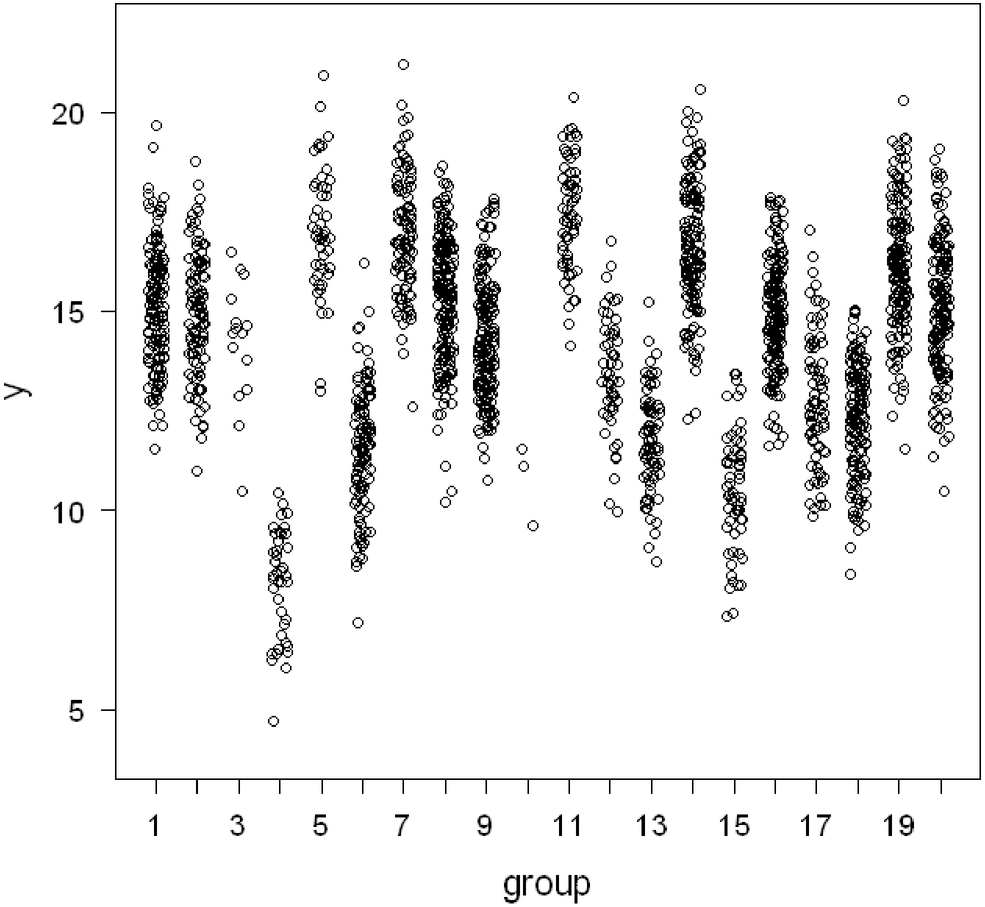
fixed effects	**random effects**
the levels (not the effects!) are fixed	the levels are not fixed
differences between specific levels are of interest	variance between the levels is of interest (or needs to be accounted for in the model)

Time for an exercise

(1) Take one or more examples of your own study and discuss which of the variables may be treated as „random“ and which ones as „fixed“.

Complete, partial and no pooling

20 groups with n between 1 and 200 observations and different means

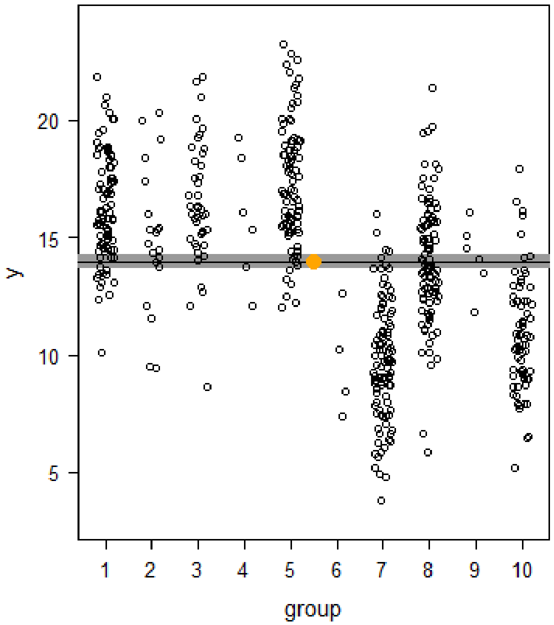


Complete, partial and no pooling

complete pooling
overall mean

$$\hat{y}_i = \beta_o$$

$$y_i \sim Norm(\hat{y}_i, \sigma^2)$$



partial pooling
mixed model

$$\hat{y}_i = \beta_o + b_{g_i}$$

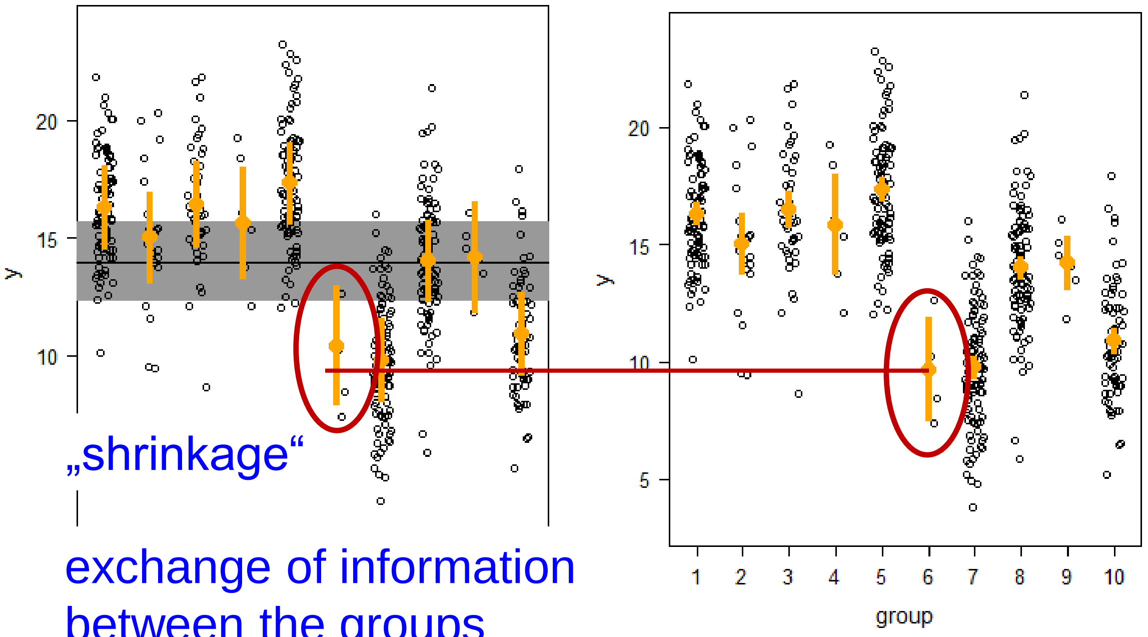
$$y_i \sim Norm(\hat{y}_i, \sigma^2)$$

$$b_g \sim Norm(0, \sigma_g^2)$$

no pooling
groupwise mean

$$\hat{y}_i = \beta_{g_i}$$

$$y_i \sim Norm(\hat{y}_i, \sigma_{g_i}{}^2)$$



„shrinkage“

exchange of information
between the groups

Complete, partial and no pooling

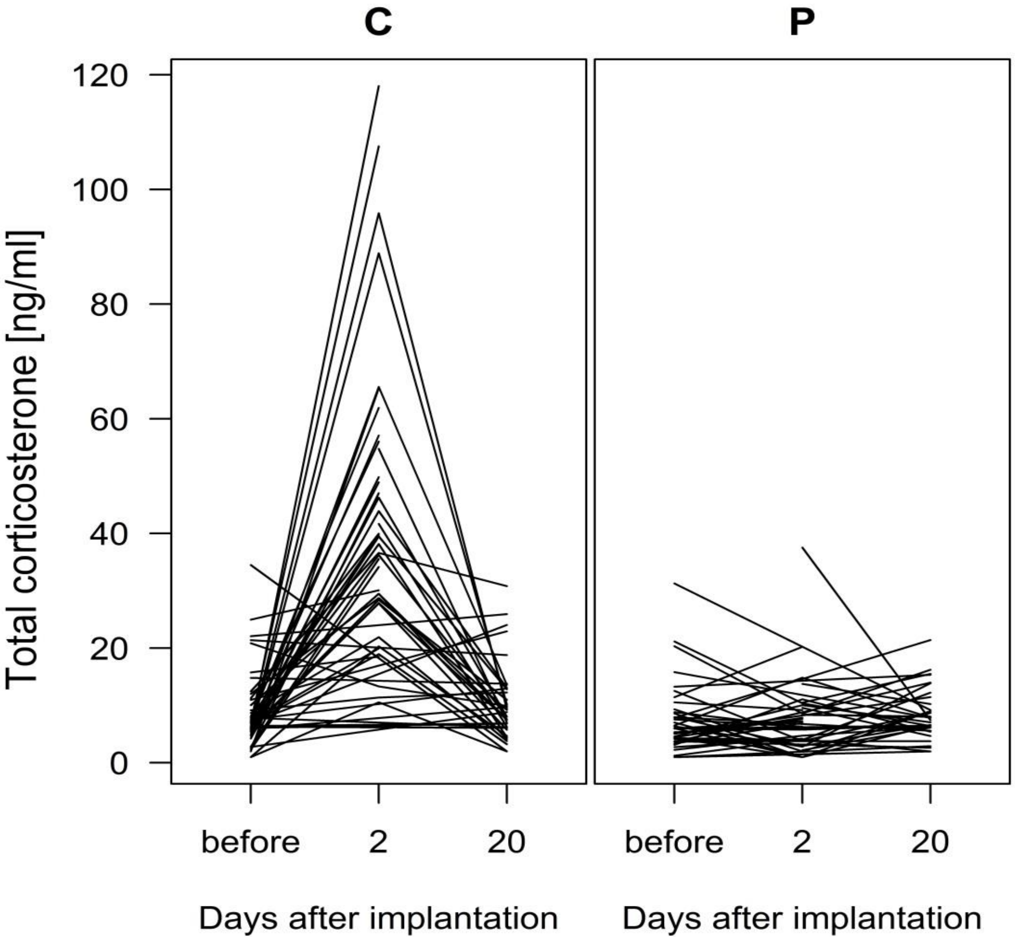
mixed model = partial pooling

group means of a mixed model = weighted averages of the overall mean and the independent group means

weights ~ sample size and between-group/within-group variance

Fitting a linear mixed model

Barn owls nestlings and stress hormone:
How does the **implant** affect **corticosterone concentration**?



Fitting a linear mixed model

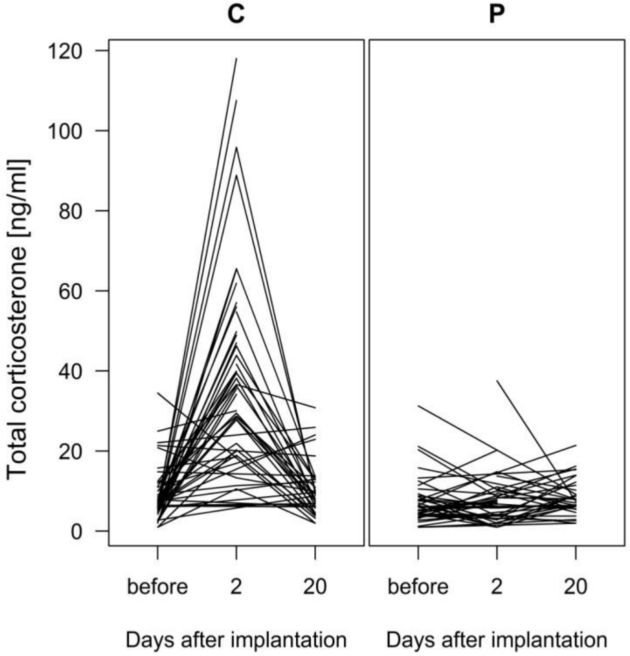
Barn owls nestlings and stress hormone:
How does the implant affect corticosterone concentration?

$$\hat{y}_i = \beta_o + b_{Ring_i} + \beta_1 I(Implant_i = P) + \beta_2 I(days_i = 3) + \beta_3 I(days_i = 21) +$$

$$\beta_4 I(Implant_i = P) I(days_i = 3) + \beta_5 I(Implant_i = P) I(days_i = 21)$$

$$y_i \sim Norm(\hat{y}_i, \sigma^2)$$

$$b_{Ring} \sim Norm(0, \sigma_b^2)$$



Fitting a linear mixed model

```
library(lme4)

mod.REML=lmer(log(totCort) ~Implant + days

                + Implant:days

                + (1|Ring), data=dat, REML=TRUE)
```

Fixed effects

Interaction

random effect

Fitting a linear mixed model

```
mod.REML=lmer(log(totCort) ~Implant + days+ Implant:days +
 (1|Ring), data=dat, REML=TRUE)

summary(mod.REML)
```

restricted maximum likelihood

```
mod.REML
Linear mixed model fit by REML ['lmerMod']
Formula: log(totCort) ~ Implant + days + Implant:days + (1 | Ring)
   Data: dat
REML criterion at convergence: 611.9053
Random effects:
 Groups   Name        Std.Dev.
 Ring     (Intercept) 0.3384
 Residual             0.6134
Number of obs: 287, groups:  Ring, 151
Fixed Effects:
(Intercept) ImplantP   days2  days20 ImplantP:days2 ImplantP:days20
    1.91446 -0.08523 1.65307 0.26278       -1.71999        -0.09514
```

between-individual variance

residual variance

fixed effects

Fitting a linear mixed model

fixef(mod)

```
(Intercept) ImplantP   days2   days20   ImplantP:days2   ImplantP:days20
1.914       -0.084     1.653   0.262    -1.720           -0.095
```

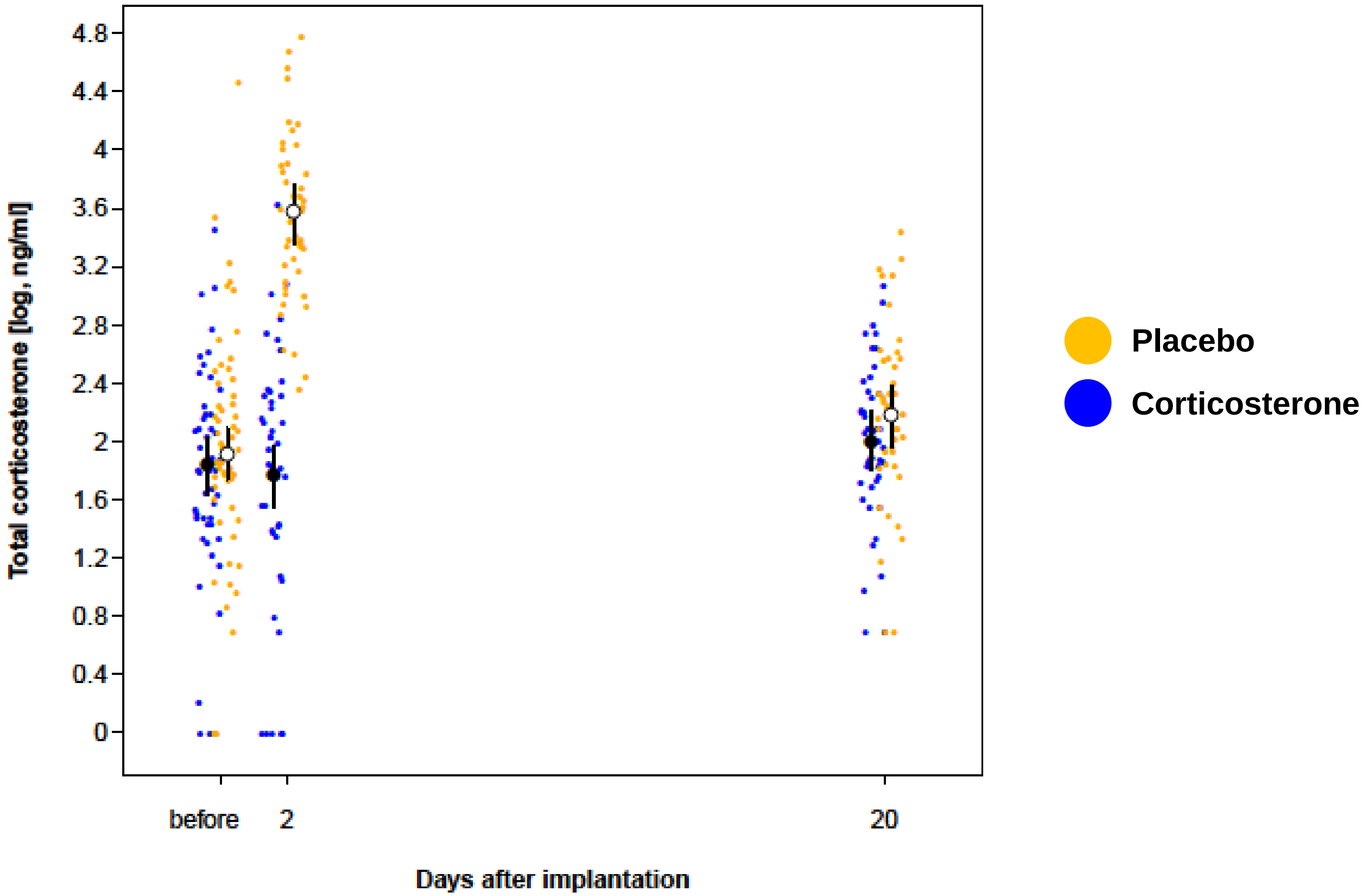
ranef(mod)

```
$Ring       Intercept)
898054    0.250867170
898055    0.119193973
898057   -0.108778053
898058    0.070320591
898059   -0.081330604
```

coef(mod)

```
$Ring     (Intercept)    ImplantP    days2   days20 ImplantP:days2 ImplantP:days20
898054      2.165247 -0.08488857 1.653434 0.262791      -1.720644     -0.09530968
898055      2.033574 -0.08488857 1.653434 0.262791      -1.720644     -0.09530968
898057      1.805602 -0.08488857 1.653434 0.262791      -1.720644     -0.09530968
898058      1.984700 -0.08488857 1.653434 0.262791      -1.720644     -0.09530968
898059      1.833049 -0.08488857 1.653434 0.262791      -1.720644     -0.09530968
```

Fitting a linear mixed model



REML vs LM

Important: mixed models can be fit by ML or by REML

REML gives unbiased estimates for the variance components
but are biased in the variance of the fixed effects

Guideline in fitting mixed models:

1.use REML to analyse the random model structure

2. switch to ML to draw inference about the fixed effect part

```
mod.LM=lmer(log(totCort)~Implant + days + Implant:days
            +(1|Ring), data=dat, REML=FALSE)
```

Assesment of model assumptions

```
# residuals vs. fitted (homogeneity of variance and distribution)
scatter.smooth(fitted(mod),resid(mod)); abline(h=0, lty=2)
title("Tukey-Anscombe Plot")
```

→ **no pattern**

```
# qq of residuals (normality)
qqnorm(resid(mod), main="normal QQ-plot, residuals")
qqline(resid(mod))
```

→ **no deviation from the main line**

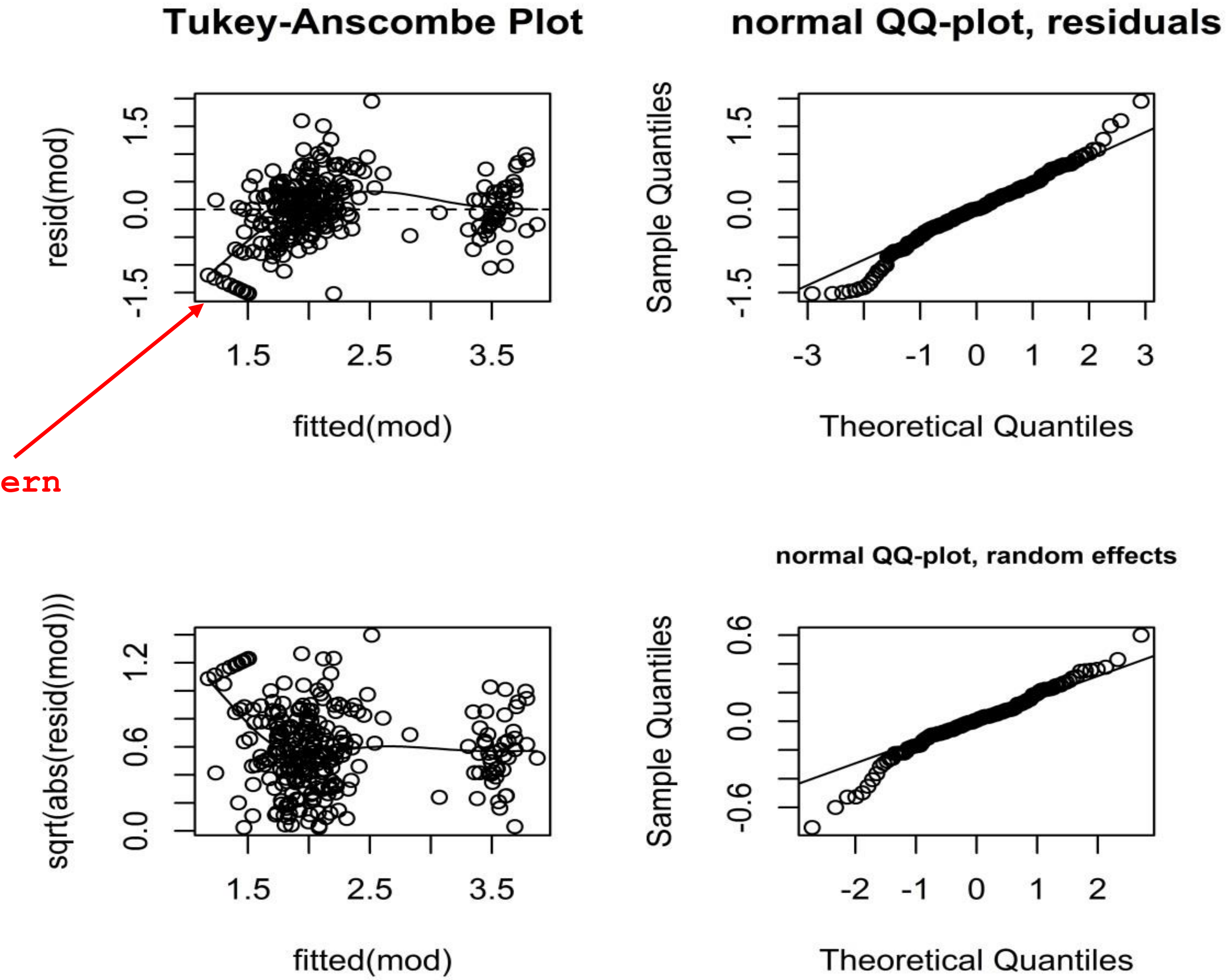
```
# (squarerooted) residuals vs. fitted (homogeneity of variance )
scatter.smooth(fitted(mod), sqrt(abs(resid(mod))))
```

→ **no pattern**

```
# qq of random effects (normality of random effects)
qqnorm(ranef(mod)$Ring[,1])
qqline(ranef(mod)$Ring[,1])
```

→ **no deviation from the main line**

Assesment of model assumptions



Tukey-Anscombe Plot

normal QQ-plot, residuals

normal QQ-plot, random effects

pattern

Time for an exercise

Use the dataset “Forest_data.csv”. Data on individual trees taken on plots along transects in a few sections of a valley in Washington state where trees were encroaching into shrub steppe habitat .

(1) Input the data into R

(2) Perform data exploration

(3) Does **tree height** within each plot depend on the **slope** of the plot? Fit the appropriate model. What do you need to take into account? What could be a problem?

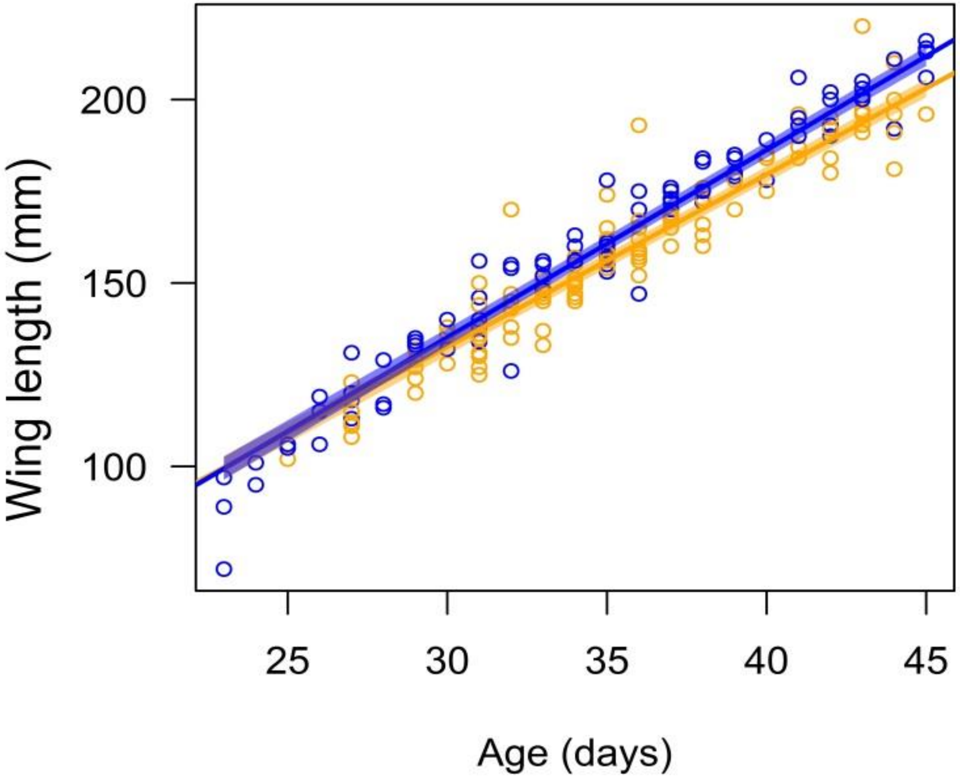
(4) Perform model validation.

(5) Think about a question where the random factors you used in (3) become fixed factors.

Random intercepts and slope

Barn owls nestlings and stress hormone:
How does **corticosterone** treatement affect **growth rate**?

→ Besides having a different intercept, different individuals are also likely to react slightly differently to the corticosterone treatment



● Placebo

● Corticosterone

Random intercepts and slope

basic model with interaction:

$$\hat{y}_i = \beta_o + \beta_1 age_i + \beta_2 I(Implant = P) + \beta_3 age_i I(Implant = P)$$
$$y_i \sim Norm(\hat{y}_i, \sigma^2)$$

plus between-individual variance in growth rate

$$\hat{y}_i = \beta_o + \boxed{b_{1,Ring_i}} + (\beta_1 + \boxed{b_{2,Ring_i}})age_i + \beta_2 I(Implant = P) + \beta_3 age_i I(Implant = P)$$
$$y_i \sim Norm(\hat{y}_i, \sigma^2)$$
$$\boxed{b_{1:2,Ring} \sim MVNorm(\mathbf{0}, \Sigma)}$$

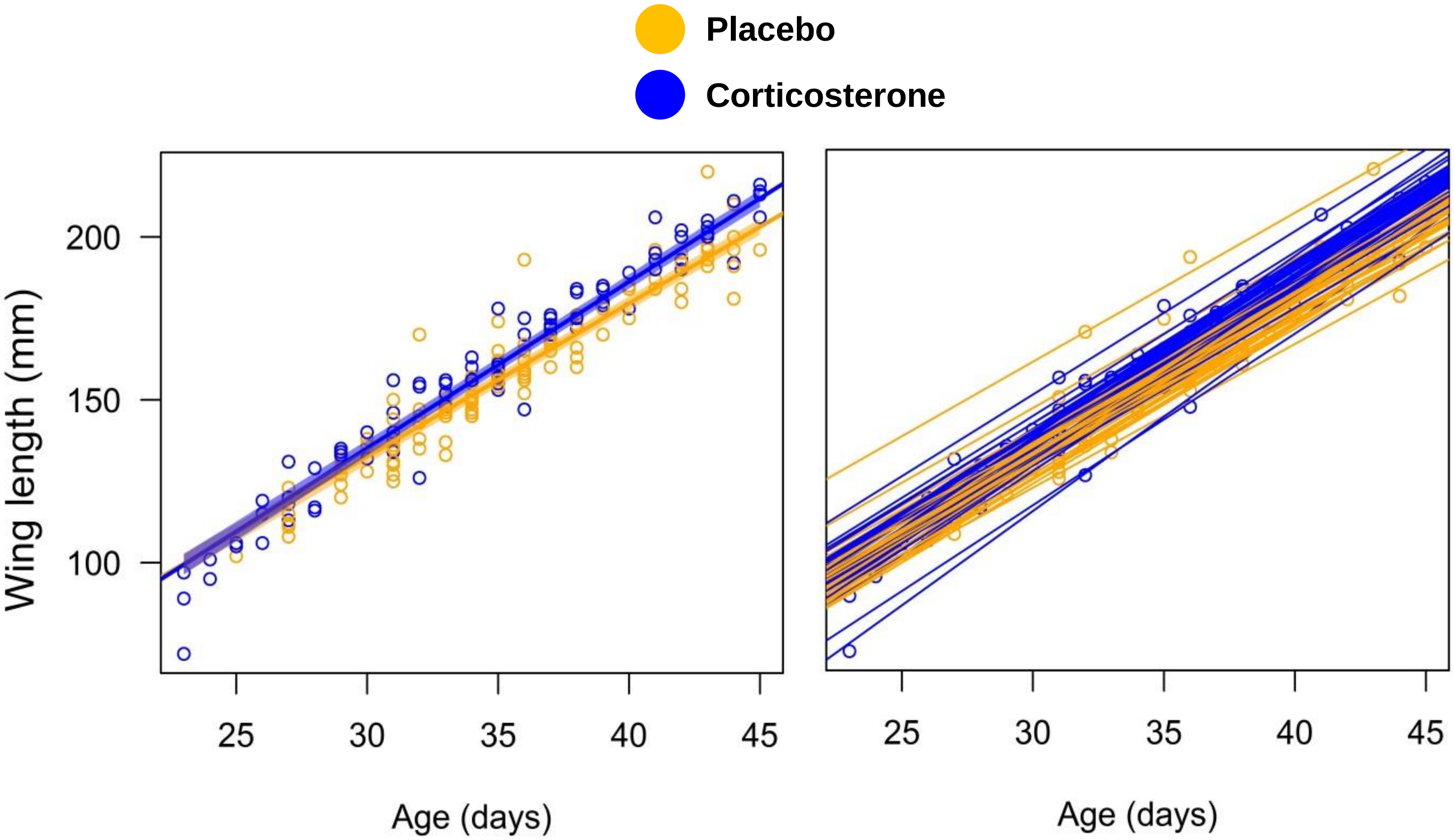
Random intercepts and slope

```
mod = lmer(Wing ~ Age + Implant + Age:Implant
                    + (Age|Ring),data=dat, REML=FALSE)
mod
```

```
Linear mixed model fit by maximum likelihood ['lmerMod']
Formula: Wing ~ Age + Implant + Age:Implant + (Age | Ring)
   Data: dat
      AIC       BIC    logLik  deviance
1280.4391 1307.1778 -632.2195 1264.4391
Random effects:
 Groups   Name        Std.Dev. Corr
 Ring     (Intercept) 14.6424
          Age          0.3573  -0.90
 Residual              2.5419
Number of obs: 209, groups: Ring, 86
Fixed Effects:
 (Intercept)          Age      ImplantP  Age:ImplantP
     -8.2887       4.6979       -9.7816        0.4113
```

Random intercepts and slope

Barn owls nestlings and stress hormone:
How does **corticosterone** treatement affect **growth rate**?



Random intercepts and slope

for the corticosterone-treated individuals:

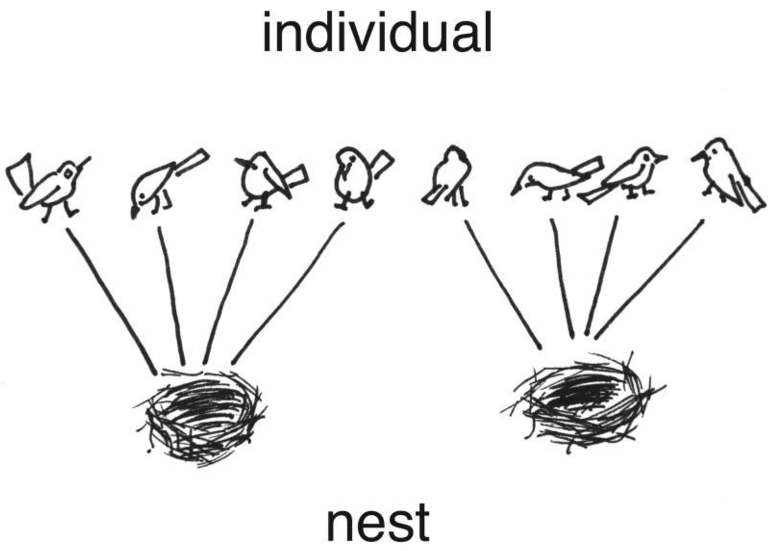
```
abline(fixef(mod)[1] +                 ranef(mod)$Ring[i,1],
       fixef(mod)[2] +                 ranef(mod)$Ring[i,2],
       col="orange")
```

for the placebo-treated individuals:

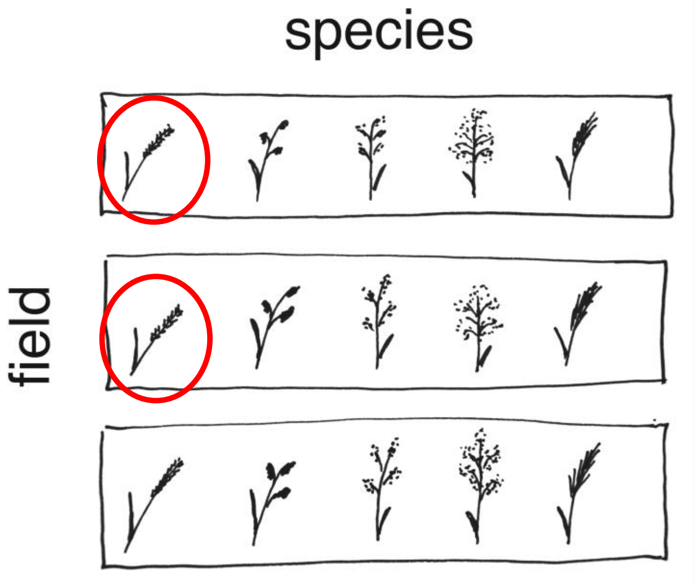
```
abline(fixef(mod)[1] + fixef(mod)[3] + ranef(mod)$Ring[i,1],
       fixef(mod)[2] + fixef(mod)[4] + ranef(mod)$Ring[i,2],
       col="blue")
```

Nested and crossed random effects

nested



crossed



Nested and crossed random effects

nested

```
mod <- lmer(log(totCort) ~ Implant + days + Implant:days +
(1|Brood/Ring), data=cortbowl, REML=FALSE)
```

crossed

```
mod <- lmer(log(totCort) ~ Implant + days + Implant:days +
(1|Brood) + (1|Ring), data=cortbowl, REML=FALSE)
```

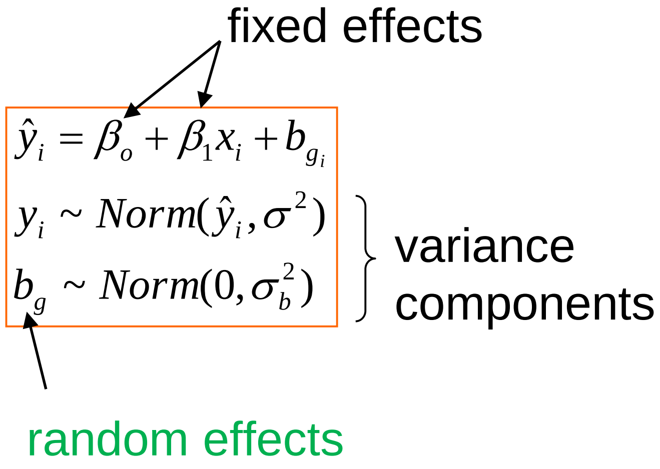
What we have learned

► use REML to analyse random effects, use ML to analyse fixed effects

► mixed models = partial pooling

► crossed random effects: y~x+(1|group1) + (1|group2)

► nested random effects: y~x+(1|group2/group1)



fixed effects

$$\boxed{\begin{aligned} &\hat{y}_i = \beta_o + \beta_1 x_i + b_{g_i} \\ &y_i \sim Norm(\hat{y}_i, \sigma^2) \\ &b_g \sim Norm(0, \sigma_b^2) \end{aligned}}$$

variance components

random effects

Time for an exercise

Use the dataset “Fields.txt”. The dataset contains the number of plant species and soil nitrate levels in 126 plots of vegetation.

(1) Input the data into R

(2) Perform data exploration

(3) Fit a model to look at the relationship between **soil nitrate level** and **species richness**. What do you need to take into account?

(4) Perform model validation. What do you notice? What is the problem? How can it be solved?

Practice at home

Follow up on the cortbowl.txt dataset we used for the nested model example.

(1) Draw a graph of the fitted model, specifying a line for each individual.

(2) Draw a graph of the fitted model, specifying a line for each brood.

Acknowledgements

People:

Noelie Maurel
Wayne Dawson
Fränzi Körner

International Max Planck
Research School
for Organismal Biology

Ecology
Konstanz

Institute for
Health Metrics
and Evaluation

TReND

The Company of
Biologists

Supported by

The Company of Biologists

Development

Journal of Cell Science

Journal of Experimental Biology

Disease Models & Mechanisms

Biology Open