

MIXED EFFECT MODELS

PSEUDO-REPLICATION AND HETEROGENEITY

Alejandro Ordonez

Assistant Professor - Department of Bioscience

Section for Ecoinformatics & Biodiversity

Center for Biodiversity Dynamics in a Changing World (BIOCHANGE)

MIXED EFFECT MODELS

WHAT WE WILL TALK ABOUT TODAY

The basics of mixed effect models – the mathematical principle

Pseudo-replication – Solving non independence

- Nested designs and Random Effects

Model selection – Combing ML and RMEL

Any questions?

Ready to start?

MIXED EFFECTS MODELS

THE PROBLEM OF INDEPENDENCE

Observational/experiments can have:

- A nested observational/experimental design
 - samples taken across different spatial-temporal-taxonomic scales
- Spatial-Temporal-Evolutionary proximity
 - Autocorrelation → removes independence between samples

What does this mean?
Samples have “Pseudoreplication”

MIXED EFFECTS MODELS

(PSEUDO)REPLICATION

What is replication?

Independent observations of same measurement

What is
(Pseudo)replication?

Replicates are not independent (spatial-temporal), or
“treatments are truly not replicated”

What is independent?

Independent is when the value of one observation does
not affect the value of other observations

INDEPENDENCE OF OBSERVATIONS

What does it mean?

- Objects are sampled only once.
- The occurrence of one event doesn't change the probability for another.
- Sampling of one observation does not affect the choice of the second observation.

How can we ensure independence of observations?

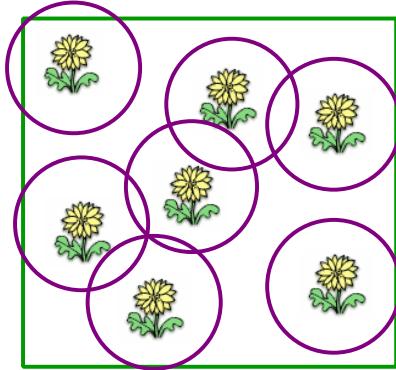
- Sample each object only once.
- Make sure that replicates have no relation → if they do, they are repetitions.
- Use a fully randomized design !!!

(PSEUDO)REPLICATION

Probably no other aspect of experimental design causes more problems for biologists”

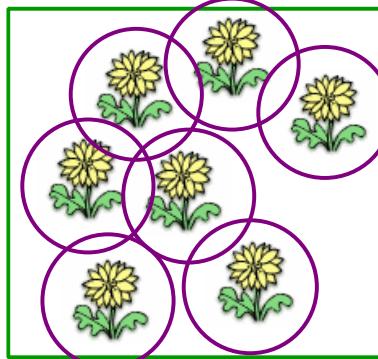
Quinn & Keough (2002) (Experimental Design & Data Analysis for Biologists)

Plot 1 (fertilizer 1)



Which is the unit of replication in the figure? → Plant or Plot?

Plot 2 (fertilizer 2)



Plant is unit of replication

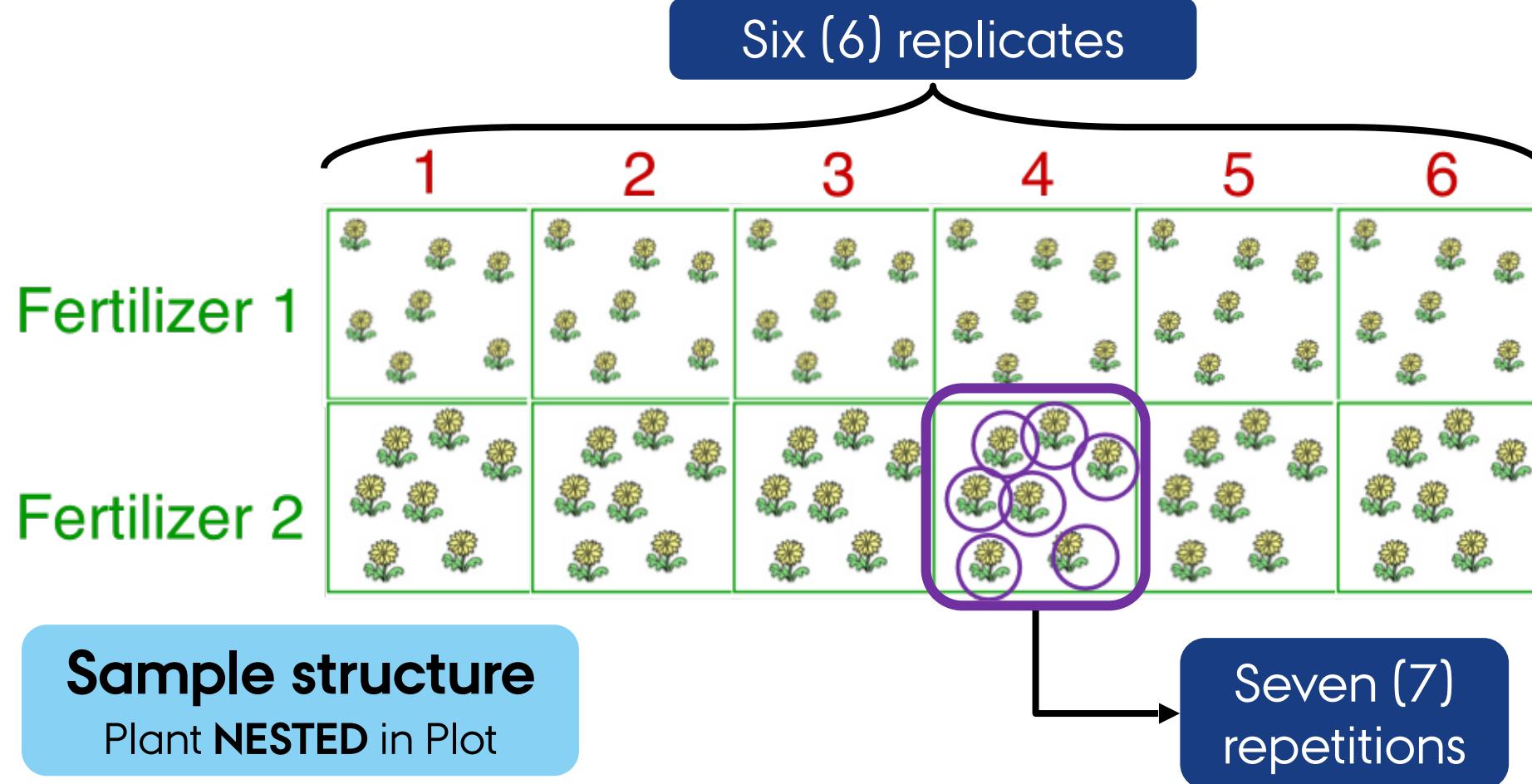
2-sample t-test /w 7 replicates
(error df=12)



Plot is unit of replication

2-sample t-test /w 1 replicate
(error df=0)

(PSEUDO)REPLICATION



ADDRESSING (PSEUDO)REPLICATION

TWO ALTERNATIVE GOALS

Management of experimental error

- Where does most of the variation occur?
- Where would increased replication be most profitable?

In a nutshell:

Management of experimental error is all about degrees of freedom budgets

Partition the variation across different levels/scales

- At which level/scale is most of the variation generated ?

In a nutshell:

Partition the variation goal is to describe how each level/scale determines the population variance

So far so good?

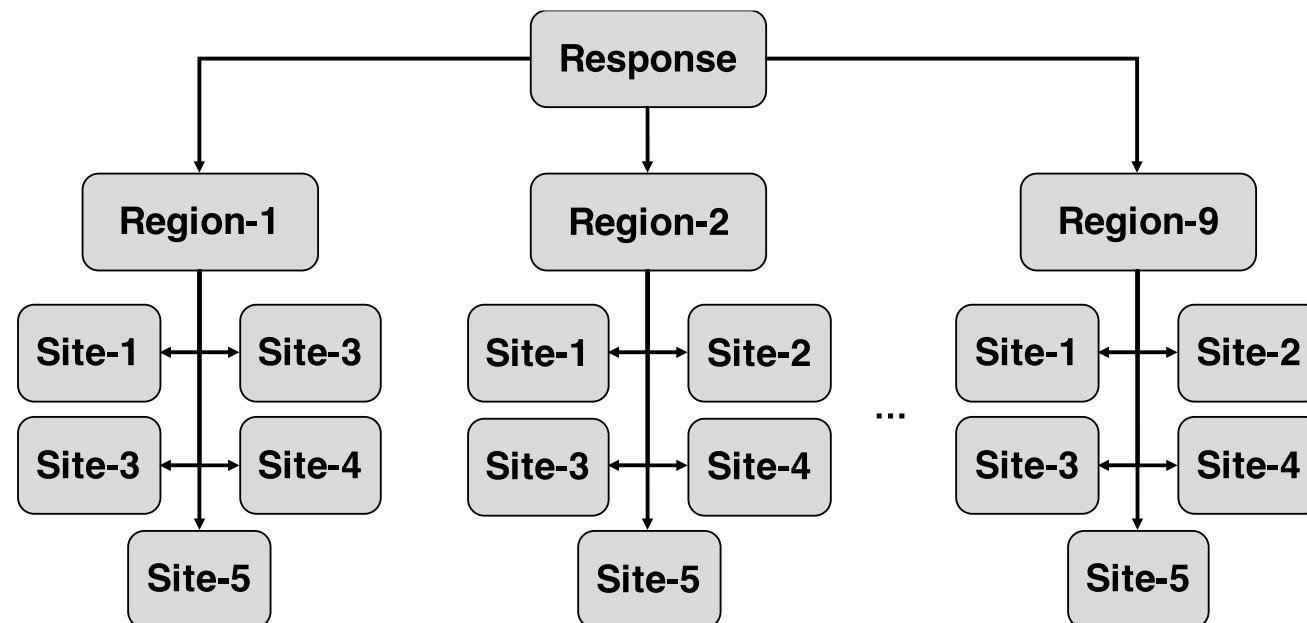
Any questions?

Ready to move on?

NESTED DESIGNS & (PSEUDO)REPLICATION

What nested design means?

The response varies within more than one level (also known as **hierarchical design**)

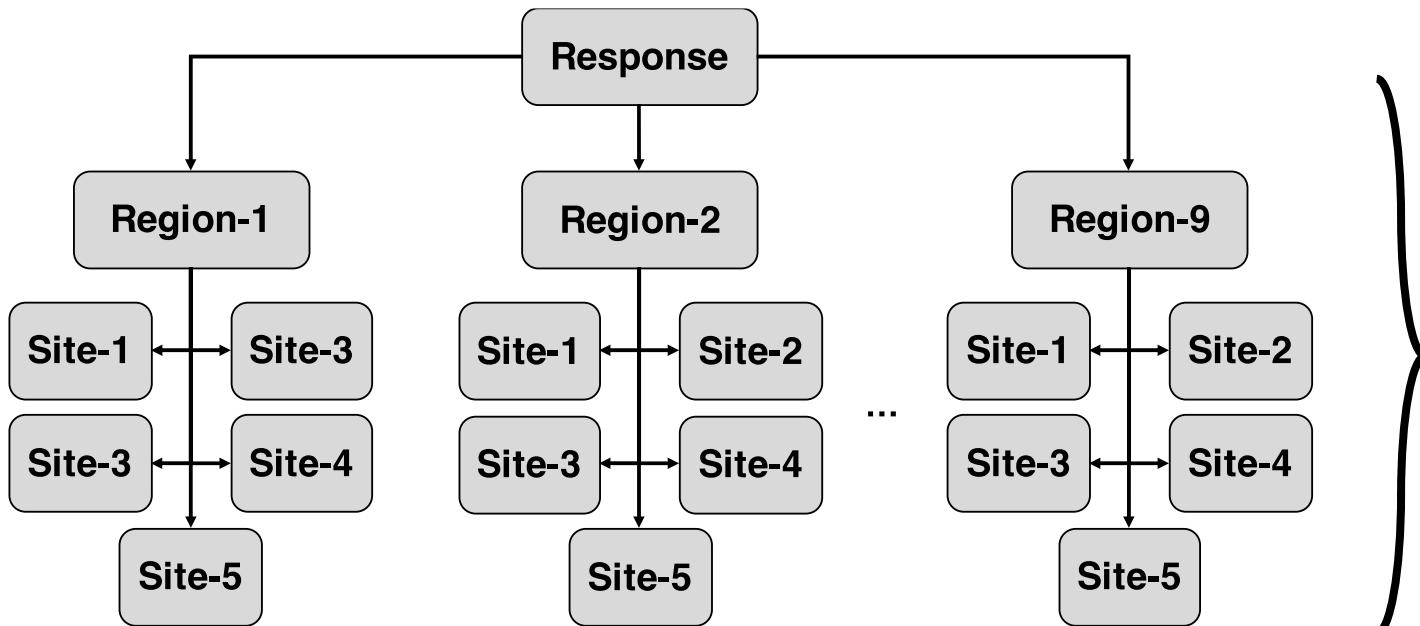


Observations
(sites) within
regions are not
independent

NESTED DESIGNS & (PSEUDO)REPLICATION

Question

What is the effect on the response of changes in X?

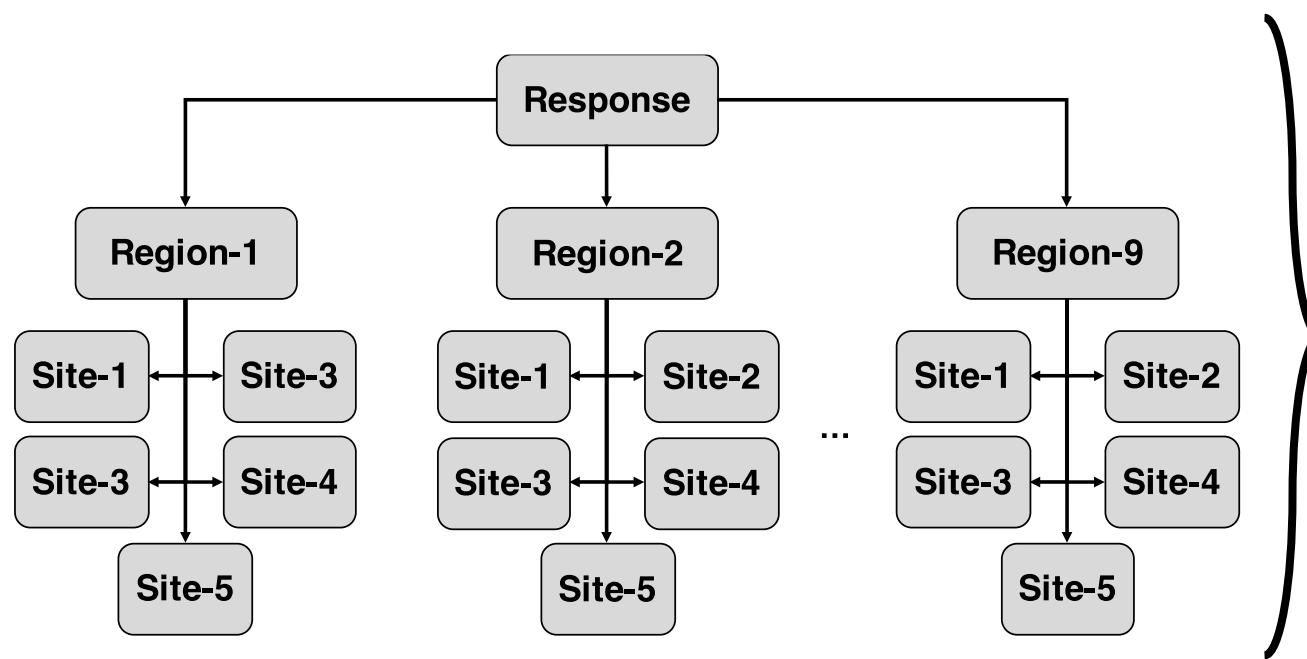


But...

We know that the effect of X
changes between nine (9)
regions

$$Y_i \sim \text{Region}_{i,1} + (X_{i,n} * \text{Region}_{i,n})$$

NESTED DESIGNS & (PSEUDO)REPLICATION



- This model needs nine (9) parameters to be estimated
- We loose **9 degrees of freedom**
 - Losing df → Reduces of inference power!

But in most cases, we don't care about the region effect

How can we deal with this?

Mixed Effect Models

Allow controlling for regional effects and not loose DF

So far so good?

Any questions?

Ready to move on?

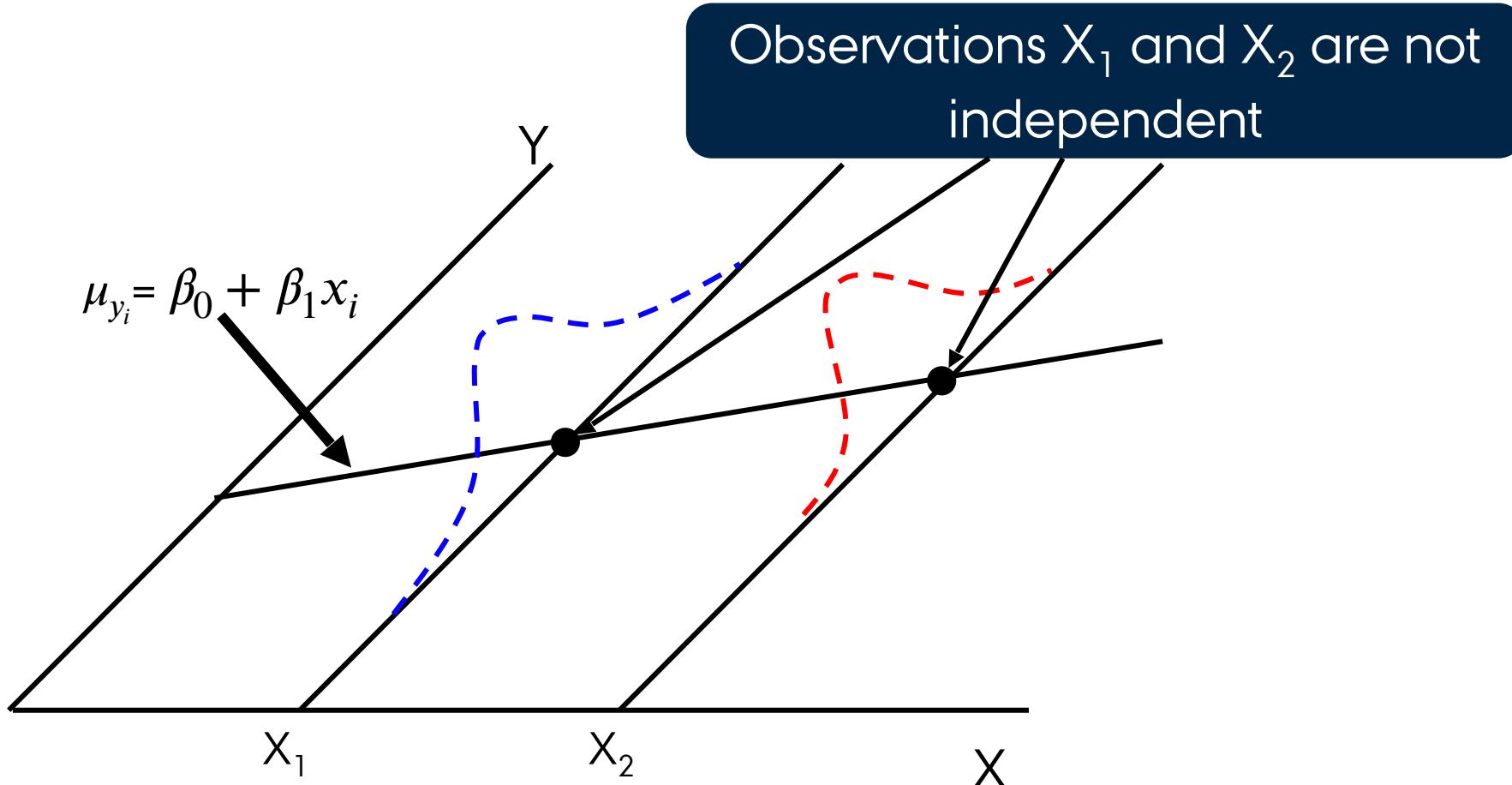
GENERAL LINEAR MODELS

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

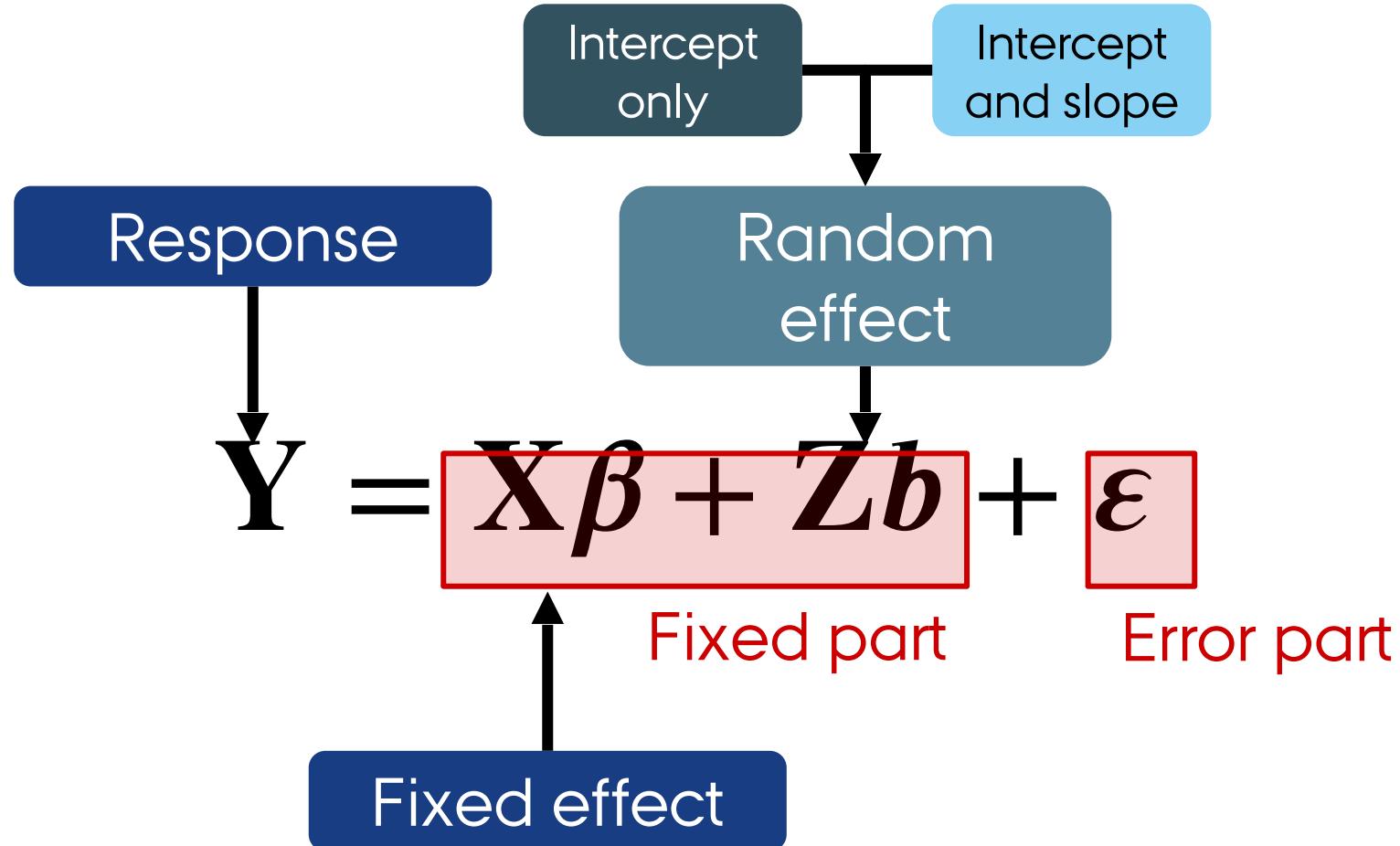
$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

The diagram illustrates the General Linear Model equation $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. The response vector \mathbf{Y} is shown as a column of y_1, \dots, y_n . The design matrix \mathbf{X} is shown as a matrix of $1, x_{ij}, \dots$ where x_{ij} represents the value of the j -th predictor variable for the i -th observation. The parameter vector $\boldsymbol{\beta}$ is shown as a column of $\beta_0, \beta_1, \beta_2$. The error term $\boldsymbol{\varepsilon}$ is shown as a column of $\varepsilon_1, \dots, \varepsilon_n$. A blue box highlights the first row of the design matrix, labeled "Fixed part". A red box highlights the error term $\boldsymbol{\varepsilon}$.

GENERAL LINEAR MODEL VIOLATING THE ASSUMPTIONS



MIXED EFFECTS MODELS



MIXED EFFECTS MODELS

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} 1 & z_{11} \\ \vdots & \vdots \\ 1 & z_{n1} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Fixed part

Error part

MIXED EFFECTS MODELS

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} 1 & z_{11} \\ \vdots & \vdots \\ 1 & z_{n1} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

This is how we deal with **Pseudoreplication** in the experimental design

Fixed effect

Random effect

Error part

MIXED EFFECTS MODELS

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} +$$
$$\begin{bmatrix} 1 & z_{11} \\ \vdots & \vdots \\ 1 & z_{n1} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} +$$
$$\begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Fixed effect

Fixed effects
influence only the
mean of y

Random effect

Random effects
influence only the
variance of y

Error part

MIXED EFFECTS MODELS

FIXED VS RANDOM EFFECTS

Fixed effects

Levels decided (fixed) by experimenter

If experiment repeated, same levels used again

Not a random sample

Cannot extrapolate to other levels

Random effects

Levels are a random selection of all possible levels

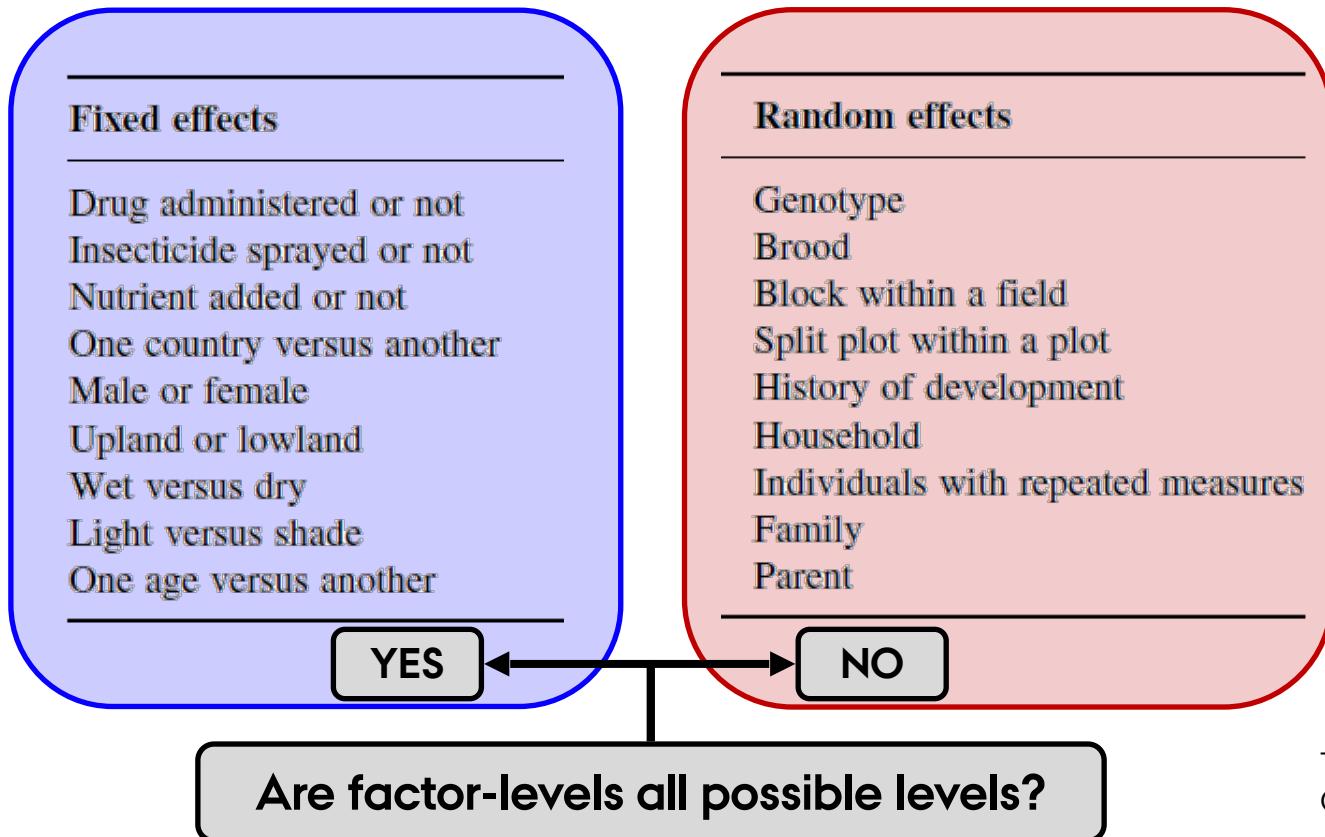
If experiment repeated, different random selection used

Useful to generalise inferences to all possible levels

MIXED EFFECTS MODELS

FIXED VS RANDOM EFFECTS

Fixed effects have
informative factor
levels



Random have
uninformative
factor levels

Table from Crawley (the R-book)
Chp 19, pp 627-630

So far so good?

Any questions?

Ready to move on?

BUILDING MIXED EFFECTS MODELS IN R

Random intercept model

Only intercepts are allowed to vary
Slopes are fixed à the same across different contexts

The mean response is believed to change among groups

Provides information about intra-class correlations

Random slope and intercept model

Both intercepts and the slopes are allowed to vary

Controls for the reduction in Degrees of Freedom

Both the mean response and the rate of change is believed to change among groups

BUILDING MIXED EFFECTS MODELS IN R

Random effect model

The variation in the response variable is only allowed to be described by random effects.

There are NO FIXED EFFECTS!!

Assesses the how the data varies based on the sampling hierarchy à Variance components

Marginal models

The within hierarchy correlation is modelled within the random part of the model

Need to define a specific correlation structure

Observations within a level are correlated

BUILDING MIXED EFFECTS MODELS IN R

Random intercept model

```
lme(lme(Y~Fix.Eff,  
        random=~1|Rand.Eff,  
        data=data.In))
```

Random slope and intercept model

```
lme(Y~Fix.Eff,  
    random=~1+Fix.Eff|Rand.Eff,  
    data=data.In)
```

Random effect model

```
lme(lme(Y~1,  
        random=~1|Rand.Eff,  
        data=data.In))
```

Marginal models

```
gls(Y~Fix.Eff,  
correlation=corCompSym(~1|Rand.Eff),  
data=data.In)
```

So far so good?

Any questions?

Ready to move on?

MIXED EFFECTS MODELS

THE CORE ASSUMPTIONS

1. Within-group errors are independent with mean zero and variance σ^2 .
2. Within-group errors are independent of the random effects.
3. The random effects are
 - Normally distributed.
 - With mean zero and equal covariance matrix → for the linear case.
4. The random effects are independent in different groups.
5. The covariance matrix does not depend on the group.

So far so good?

Any questions?

Ready to move on?

BUILDING MIXED EFFECTS MODELS IN R

TWO APPROACHES

Random intercept

Only intercepts are allowed to vary.

```
lme(Y~Fix.Eff,  
     random=~1|Rand.Eff,  
     data=data.In)
```

ONLY the mean response is believed to change among groups

Random slope-intercept

Intercepts **AND** the slopes are allowed to vary.

```
lme(Y~Fix.Eff,  
     random=~1+Fix.Eff|Rand.Eff,  
     data=data.In)
```

Both the mean response and the rate of change is believed to change among groups

MIXED EFFECTS MODELS

TWO DIFFERENT PACKAGES

Linear and Nonlinear Mixed Effects Models

- nlme package → The initial package
- Best suited for nested/repeated measures designs but...

Only for Gaussian models

- Easy to specify variance-covariance structures for the residuals when you have:
 - Heteroskedasticity
 - Covariate-dependent variability (nested)
 - Spatial/Temporal autocorrelation

MIXED EFFECTS MODELS

TWO DIFFERENT PACKAGES

Linear Mixed-Effects Models using Eigen & S4

- lme4 package → Lmer package
- Best suited for clustered data sets

Allows non-Gaussian models (logistic or Poisson)

- Useful for very specific covariance structures
 - Diagonal covariance structures
 - Unstructured covariance matrices
 - Mixed of these two

MIXED EFFECTS MODELS

TWO DIFFERENT SYNTAXES

nlme

R-syntax

```
lme(Y~Fixed.Eff,  
     random=~1 | Rand.Eff)  
lme(Y~Fixed.Eff,  
     random=~1+Fixed.Eff | Rand.Eff)
```

Rand.Eff → a/b/c
In this random structure c is nested within b, which in turn is nested within a

lme4

R-syntax

```
lmer(Y~Fixed.Eff +  
      (1 | Rand.Eff))  
lmer(Y~Fixed.Eff +  
      (Fixed.Eff | Rand.Eff))
```

lme()

object

an object inheriting from class `lme`, representing a fitted linear mixed-effects model.

fixed

a two-sided linear formula object describing the fixed-effects part of the model, with the response on the left of a `~` operator and the terms, separated by `+` operators, on the right, an "`lmList`" object, or a "`groupedData`" object.

There is limited support for formulae such as `resp ~ 1` and `resp ~ 0`, and less prior to version 3.1-112.

fixed.

Changes to the fixed-effects formula – see `update.formula` for details.

data

an optional data frame containing the variables named in `fixed`, `random`, `correlation`, `weights`, and `subset`. By default the variables are taken from the environment from which `lme` is called.

random

optionally, any of the following: (i) a one-sided formula of the form `~ x1 + ... + xn | g1/.../gm`, with `x1 + ... + xn` specifying the model for the random effects and `g1/.../gm` the grouping structure (`m` may be equal to 1, in which case no `/` is required). The random effects formula will be repeated for all levels of grouping, in the case of multiple levels of grouping; (ii) a list of one-sided formulas of the form `~ x1 + ... + xn | g`, with possibly different random effects models for each grouping level. The order of nesting will be assumed the same as the order of the elements in the list; (iii) a one-sided formula of the form `~ x1 + ... + xn`, or a `pdMat` object with a formula (i.e. a non-NULL value for `formula(object)`), or a list of such formulas or `pdMat` objects. In this case, the grouping structure formula will be derived from the data used to fit the linear mixed-effects model, which should inherit from class "`groupedData`"; (iv) a named list of formulas or `pdMat` objects as in (iii), with the grouping factors as names. The order of nesting will be assumed the same as the order of the elements in the list; (v) an `reStruct` object. See the documentation on `pdClasses` for a description of the available `pdMat` classes. Defaults to a formula consisting of the right hand side of `fixed`.

correlation

an optional `corStruct` object describing the within-group correlation structure. See the documentation of `corClasses` for a description of the available `corStruct` classes. Defaults to `NULL`, corresponding to no within-group correlations.

weights

an optional `varFunc` object or one-sided formula describing the within-group heteroscedasticity structure. If given as a formula, it is used as the argument to `varFixed`, corresponding to fixed variance weights. See the documentation on `varClasses` for a description of the available `varFunc` classes. Defaults to `NULL`, corresponding to homoscedastic within-group errors.

lmer()

formula

a two-sided linear formula object describing both the fixed-effects and random-effects part of the model, with the response on the left of a `-` operator and the terms, separated by `+` operators, on the right. Random-effects terms are distinguished by vertical bars (`|`) separating expressions for design matrices from grouping factors. Two vertical bars (`||`) can be used to specify multiple uncorrelated random effects for the same grouping variable. (Because of the way it is implemented, the `||`-syntax works only for design matrices containing numeric (continuous) predictors; to fit models with independent categorical effects, see [dummy](#) or the `lmer_alt` function from the `afex` package.)

data

an optional data frame containing the variables named in `formula`. By default the variables are taken from the environment from which `lmer` is called. While `data` is optional, the package authors *strongly* recommend its use, especially when later applying methods such as `update` and `drop1` to the fitted model (*such methods are not guaranteed to work properly if `data` is omitted*). If `data` is omitted, variables will be taken from the environment of `formula` (if specified as a formula) or from the parent frame (if specified as a character vector).

REML

logical scalar - Should the estimates be chosen to optimize the REML criterion (as opposed to the log-likelihood)?

control

a list (of correct class, resulting from [lmerControl\(\)](#) or [glmerControl\(\)](#) respectively) containing control parameters, including the nonlinear optimizer to be used and parameters to be passed through to the nonlinear optimizer, see the `*lmerControl` documentation for details.

start

a named [list](#) of starting values for the parameters in the model. For `lmer` this can be a numeric vector or a list with one component named "theta".

verbose

integer scalar. If `> 0` verbose output is generated during the optimization of the parameter estimates. If `> 1` verbose output is generated during the individual PIRLS steps.

subset

an optional expression indicating the subset of the rows of `data` that should be used in the fit. This can be a logical vector, or a numeric vector indicating which observation numbers are to be included, or a character vector of the row names to be included. All observations are included by default.

weights

an optional vector of 'prior weights' to be used in the fitting process. Should be `NULL` or a numeric vector. Prior weights are *not* normalized or standardized in any way. In particular, the diagonal of the residual covariance matrix is the squared residual standard deviation parameter `sigma` times the vector of inverse weights. Therefore, if the weights have relatively large magnitudes, then in order to compensate, the `sigma` parameter will also need to have a relatively large magnitude.

So far so good?

Any questions?

Ready to move on?



QUESTION TO ANSWER:

Does species richness changes as a function of exposure and NAP?

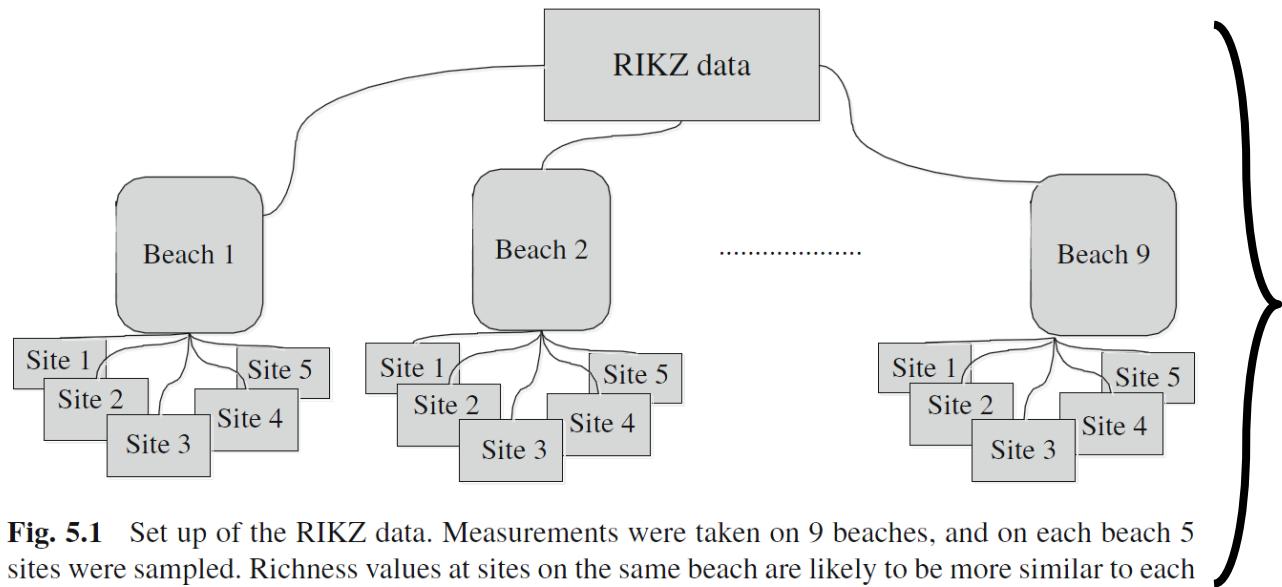
Zuur et al. (2007 - Analysing Ecological Data)
They measured benthic species richness in nine (9) inter-tidal areas along the Dutch coast. For each inter-tidal area, five (5) sites were measured.

Here we will focus on three variables:

- **Response**
 - Richness of benthic species (No of species, Richness).
- **Predictors**
 - Height of a sampling station compared to mean tidal level (in cm, NAP)
 - Exposure (a High/Low index summarising wave action, length of the surf zone, slope, grain size, and the depth of the anaerobic layer, Exposure).

MIXED EFFECTS MODELS

PRACTICAL ASPECTS (DESIGN)



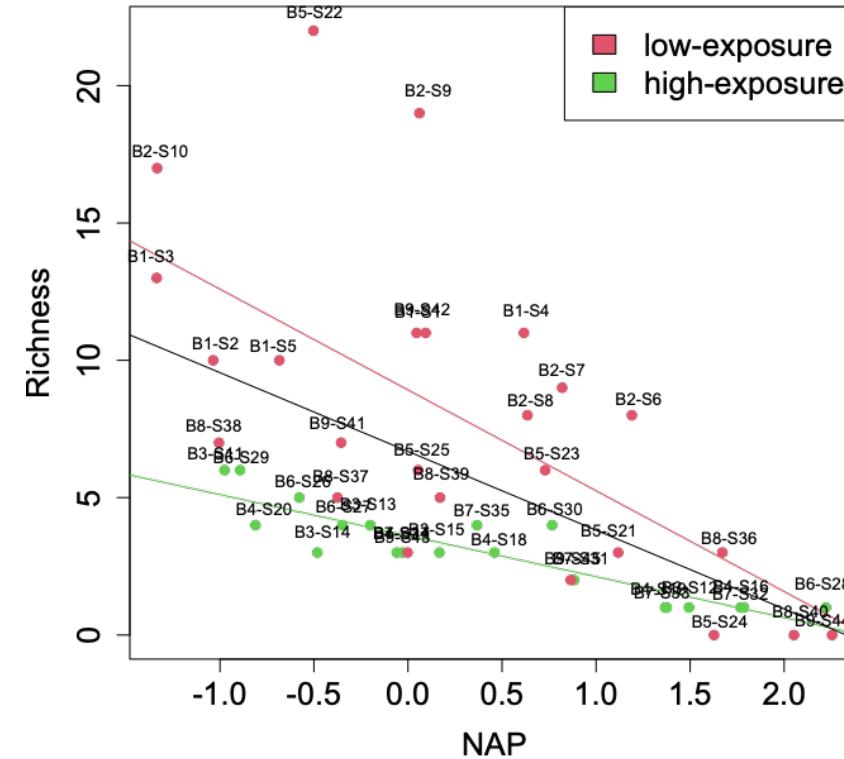
Hierarchical sampling design
Sites NESTED within Beach
Individual observations are not replicates

Fig. 5.1 Set up of the RIKZ data. Measurements were taken on 9 beaches, and on each beach 5 sites were sampled. Richness values at sites on the same beach are likely to be more similar to each other than to values from different beaches

Problem: Observations (sites) within beaches not independent
Solution: Include the sampling design as part of the model

THE PROCEDURE

1. Applying a model without a controlling for sampling design.
2. Use graphical tools to asses problems with the pseudo-replication in the sample design.
3. Build models that control for the pseudo-replication in the sample design.
4. Assess which model is the best in control for the pseudo-replication in the sample design problem.
5. Select the best combination of predictors.



Is a lm the best approach to model this relation?

Not a simple answer

Depends on → Sampling design and the structure of errors

Does Richness changes a a function of NAP and Exposure?

Yes!! Species richness decreases as a function of NAP and Exposure

```
> LM.1<-lm(Richness ~ NAP * fExp, data = RIKZ)
> summary(LM.1)

Call:
lm(formula = Richness ~ NAP * fExp, data = RIKZ)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.9289 -1.3930 -0.3001  0.9302 11.2329 

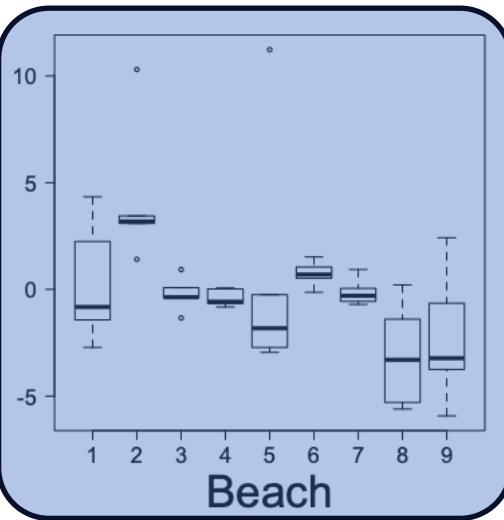
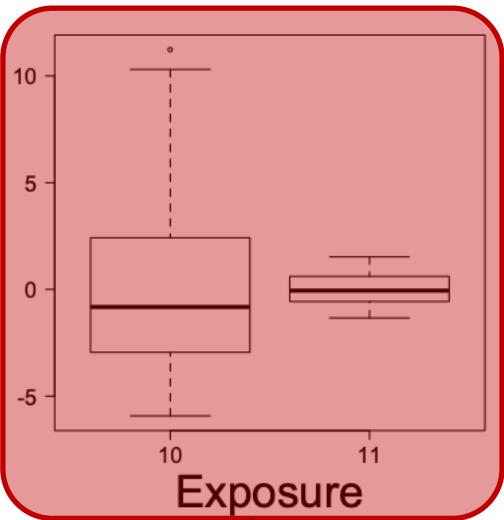
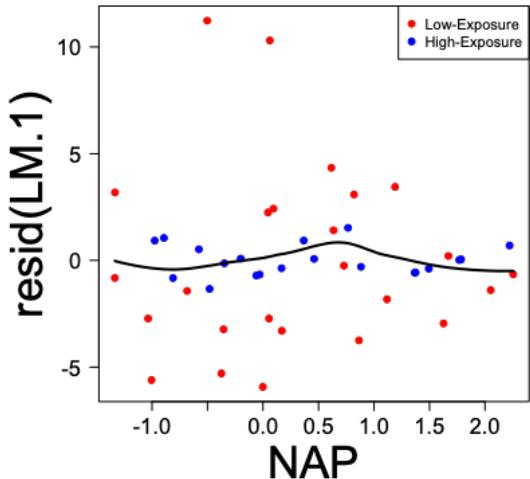
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.9216   0.7049 12.657 9.50e-16 ***
NAP        -3.6689   0.6806 -5.391 3.17e-06 ***
fExp11     -5.3043   1.0827 -4.899 1.55e-05 ***
NAP:fExp11  2.1771   1.0362  2.101  0.0418 *  
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.379 on 41 degrees of freedom
Multiple R-squared:  0.5752, Adjusted R-squared:  0.5441 
F-statistic: 18.5 on 3 and 41 DF,  p-value: 9.591e-08
```

MIXED EFFECTS MODELS

PRACTICAL ASPECTS (VARIANCE)

Is there a structure in the residuals variability?



Heteroscedasticity

Pseudo replication

MIXED EFFECTS MODELS

EQUATION NOTATION

Random Intercept model:

$$y_{ij} = \alpha + \beta \text{NAP}_{ij} + u_i + e_{ij}$$

y_{ij} species richness at beach i and site j

α fixed overall intercept

β fixed overall slope

u_i random intercept for beach i $u_i : N(0, d^2)$

e_{ij} residuals $e_{ij} : N(0, \sigma^2)$

This results in
nine (9) intercepts $[\alpha + u_{i1}]$

MIXED EFFECTS MODELS

R - IMPLEMENTATION

```
> Mlme1 <- lme(Richness ~ NAP, random = ~1 | fBeach,  
+                  data = RIKZ)  
> summary(Mlme1)  
Linear mixed-effects model fit by REML  
Data: RIKZ  
      AIC      BIC      logLik  
 247.4802 254.525 -119.7401  
Random effects:  
 Formula: ~1 | fBeach  
          (Intercept) Residual  
 StdDev:    2.944065 3.05977  
Fixed effects: Richness ~ NAP  
              Value Std.Error DF   t-value p-value  
(Intercept) 6.581893 1.0957618 35  6.006682     0  
NAP         -2.568400 0.4947246 35 -5.191574     0  
Correlation:  
  (Intr)  
NAP -0.157
```

Random Intercept
model with `lme`

$$u_i: N(0, d^2) \quad e_{ij}: N(0, \sigma^2)$$
$$\hat{d} = 2.94 \quad \hat{\sigma} = 3.06$$

MIXED EFFECTS MODELS

R - IMPLEMENTATION

```
> Mlme2 <- lmer(Richness ~ NAP + (1 | fBeach),  
+                   data = RIKZ, REML = TRUE )  
> summary(Mlme2)  
Linear mixed model fit by REML ['lmerMod']  
Formula: Richness ~ NAP + (1 | fBeach)  
Data: RIKZ  
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.  
fBeach   (Intercept) 8.668    2.944  
Residual            9.362    3.060  
Number of obs: 45, groups: fBeach, 9  
Fixed effects:  
            Estimate Std. Error t value  
(Intercept)  6.5819    1.0958   6.007  
NAP         -2.5684    0.4947  -5.192
```

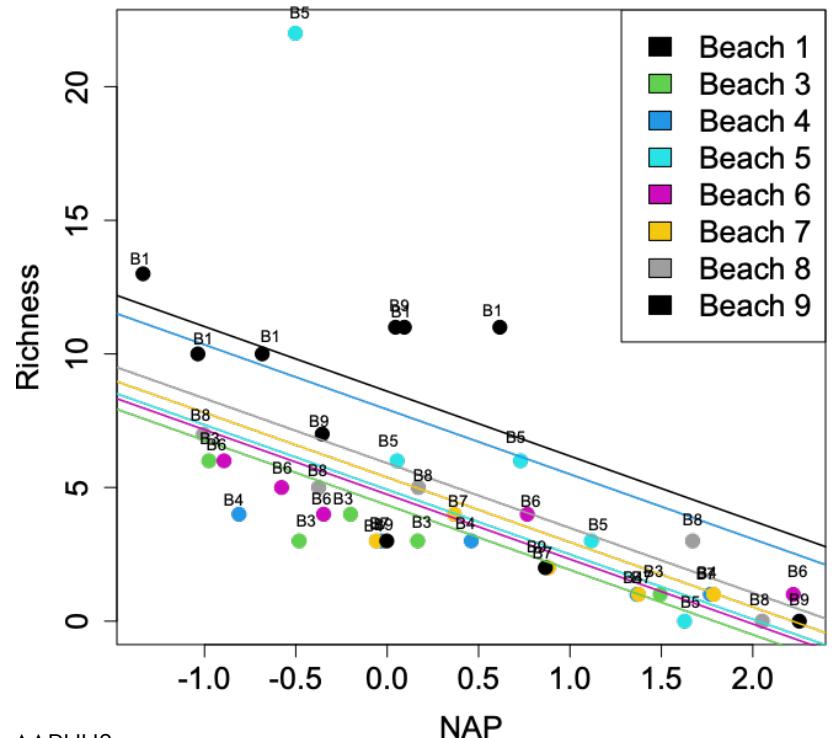
Random Intercept
Model with **lmer**

$$u_i: N(0, d^2) \quad e_{ij}: N(0, \sigma^2)$$
$$\hat{d} = 2.94 \quad \hat{\sigma} = 3.06$$

MIXED EFFECTS MODELS

R - IMPLEMENTATION

Random Intercept model:



$$y_{ij} = \alpha + \beta \text{NAP}_{ij} + u_i + e_{ij}$$

Extract the random effects from `lme`

```
> ranef(Mlme1)
  (Intercept)
1   2.621519
2   5.199608
3  -2.615780
4  -2.275618
5   1.950179
6  -1.629402
7  -1.765477
8  -1.061665
9  -0.423364
```

u_i

MIXED EFFECTS MODELS

EQUATION NOTATION

Random Intercept and slope model:

species richness at beach i and site j

fixed overall intercept

fixed overall slope

random intercept for beach i

random slope for beach i

residuals

$$y_{ij} = \alpha + \beta \text{NAP}_{ij} + u_{i1} + u_{i2} \text{NAP}_{ij} + e_{ij}$$

}

This results in

nine (9) intercepts $[\alpha + u_{i1}]$ and nine (9) slopes $[\beta + u_{i2}]$

MIXED EFFECTS MODELS

R - IMPLEMENTATION

```
> Mlmelb <- lme(Richness ~ NAP, random = ~NAP | fBeach,  
+                      data = RIKZ)  
> summary(Mlmelb)  
Linear mixed-effects model fit by REML  
Data: RIKZ  
      AIC      BIC      logLik  
 244.3839 254.9511 -116.1919  
Random effects:  
 Formula: NAP | fBeach  
 Structure: General positive-definite, Log-Cholesky  
 parametrization  
          StdDev   Corr  
(Intercept) 3.549064 (Intr)  
NAP          1.714963 -0.99  
Residual     2.702820  
Fixed effects: Richness ~ NAP  
              Value Std. Error DF t-value p-value  
(Intercept) 6.588706 1.264761 35 5.209448 0e+00  
NAP         -2.830028 0.722940 35 -3.914610 4e-04
```

Random Intercept
& slope model
with `lme`

$\hat{d}_{11}^{\wedge} = 3.55$ Intercept Var.
 $\hat{d}_{12}^{\wedge} = 1.71$ Slope Var
 $\hat{\sigma}^{\wedge} = 2.7$ Residual

MIXED EFFECTS MODELS

R - IMPLEMENTATION

```
> Mlme2b <- lmer(Richness ~ NAP + (NAP | fBeach),  
+                     data = RIKZ, REML = TRUE )  
> summary(Mlme2b)  
Linear mixed model fit by REML ['lmerMod']  
Formula: Richness ~ NAP + (NAP | fBeach)  
Data: RIKZ  
REML criterion at convergence: 232.4  
Scaled residuals:  
    Min      1Q  Median      3Q     Max  
-1.8212 -0.3410 -0.1674  0.1925  3.0397  
Random effects:  
Groups   Name        Variance Std.Dev. Corr  
fBeach   (Intercept) 12.600   3.550  
          NAP         2.942   1.715   -0.99  
Residual             7.307   2.703  
Number of obs: 45, groups: fBeach, 9  
Fixed effects:  
            Estimate Std. Error t value  
(Intercept)  6.5884    1.2649   5.208  
NAP        -2.8301    0.7231  -3.914
```

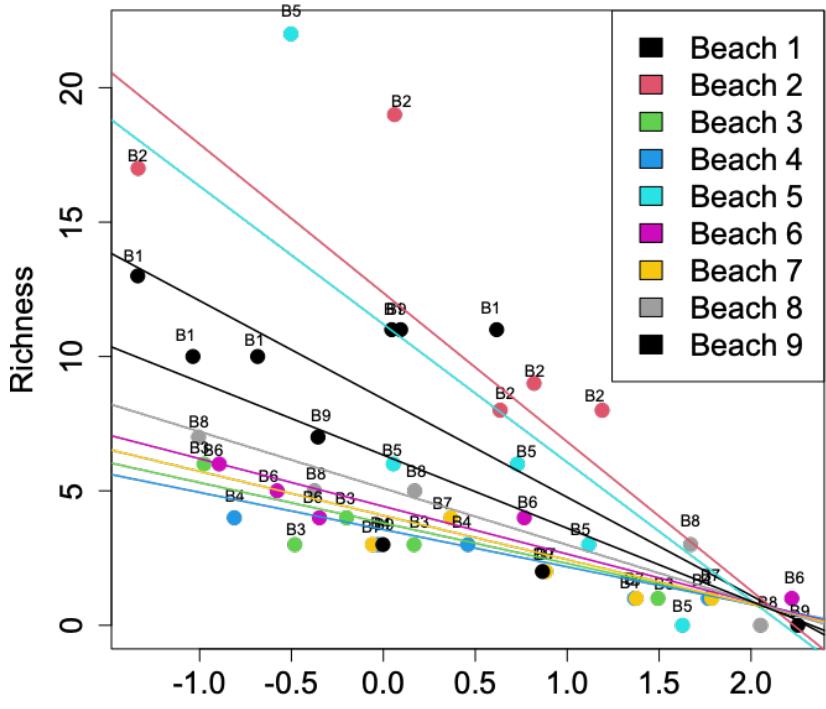
Random Intercept
& slope model with
lmer

$\hat{d}_{11} = 3.55$ Intercept Var.
 $\hat{d}_{12} = 1.71$ Slope Var
 $\hat{\sigma} = 2.703$ Residual

MIXED EFFECTS MODELS

R - IMPLEMENTATION

Random Intercept and slope model:



$$y_{ij} = \alpha + \beta \text{NAP}_{ij} + u_{i1} + u_{i2} \text{NAP}_{ij} + e_{ij}$$

Extract the random effects lme

> ranef(Mlme1b)

(Intercept)

$$u_{i1} = \left\{ \begin{array}{c} 1.8323503 \\ 5.7747909 \\ -2.7820638 \\ -3.0262848 \\ 4.6114440 \\ -2.1624275 \\ -2.5057621 \\ -1.4888117 \\ -0.2532354 \end{array} \right\} \quad \left\{ \begin{array}{c} \text{NAP} \\ -0.8262398 \\ -2.7067858 \\ 1.3243120 \\ 1.4440673 \\ -2.3072951 \\ 1.0543530 \\ 1.1856990 \\ 0.7231759 \\ 0.1087136 \end{array} \right\} = u_{i2}$$

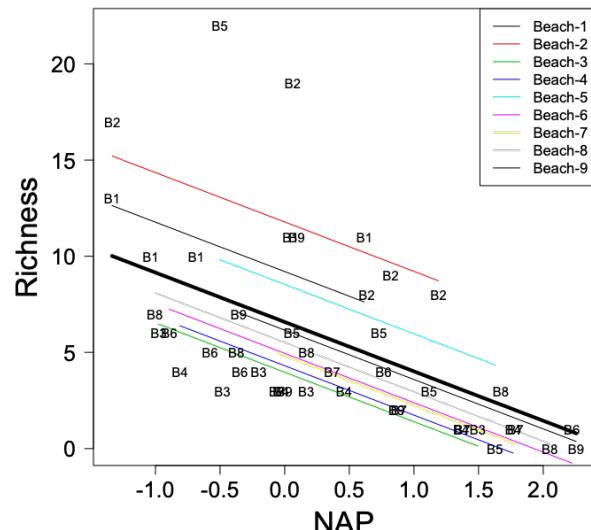
WHICH RANDOM STRUCTURE IS BETTER?

```
> anova(Mlme1, Mlme1b)
```

	Model	df	AIC	BIC	logLik	Test
Mlme1	1	4	247.4802	254.5250	-119.7401	
Mlme1b	2	6	244.3839	254.9511	-116.1919	1 vs 2

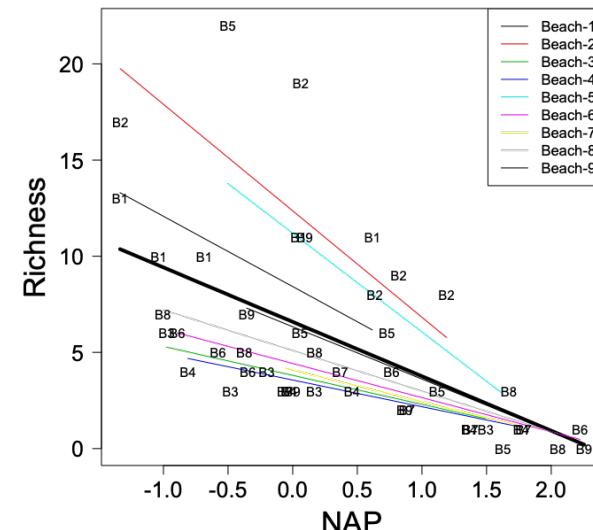
L.Ratio	p-value
7.096378	0.0288

Random Intercept
model [Mlme1]



This is a model
comparison problem
This can be done using...
AIC/BIC criteria
AND
Log-likelihood ratio test

Random Intercept/Slope
model [Mlme1b]



So far so good?

Any questions?

Ready to move on?

MIXED EFFECTS MODELS

MODEL SELECTION (PART 1)

Step -1

Fit a **FULL** model of ONLY fixed effects with the function `gls()`

This model represent of the Null hypothesis

- [H_0 = NO need for random effects]

```
B1 <- gls (Richness ~ 1 + NAP * fExp,  
            method = "REML", data = RIKZ)
```

Full = all possible fixed effects

Estimate parameters using REML

MIXED EFFECTS MODELS

MODEL SELECTION (PART 1)

Step 2

Fit FULL models of fixed effects and random effects with `lme()`

This model represent of the alternative hypothesis

- [H_A = need for random effects]

```
# Random intercept
B2 <- lme(Richness ~ 1 + NAP * fExp, data = RIKZ,
            random = ~1 | fBeach, method = "REML")
# Random slope
B3 <- lme(Richness ~ 1 + NAP * fExp, data = RIKZ,
            random = ~1 + NAP | fBeach, method = "REML")
```

Estimate
parameters
using **REML**

MIXED EFFECTS MODELS

MODEL SELECTION (PART 2)

Step 3

Select the best random structure

- Use the AIC criteria (lowest the better)

AIC (B1, B2, B3)

	df	AIC
B1	5	238.5329
B2	6	236.4925
B3	8	237.1331

→ Lowest AIC → the best model

MIXED EFFECTS MODELS

MODEL SELECTION (PART 3A)

Step 3a

Simply fixed effects by removing non-significant or marginally significant terms

```
> summary(B2)
Linear mixed-effects model fit by REML
  Data: RIKZ
        AIC      BIC      logLik
  236.4925 246.7739 -112.2462
Random effects:
  Formula: ~1 | fBeach
            (Intercept) Residual
  StdDev:     1.818579 2.942872
Fixed effects: Richness ~ 1 + NAP * fExp
                Value Std.Error DF   t-value p-value
  (Intercept) 8.861084 1.0208453 34 8.680144 0.0000
  NAP         -3.463651 0.6278583 34 -5.516613 0.0000
  fExp11      -5.255617 1.5452298  7 -3.401188 0.0114
  NAP:fExp11  2.000464 0.9461259 34  2.114374 0.0419
```

MIXED EFFECTS MODELS

MODEL SELECTION (PART 3B)

Step 3B

Test of the simple model
is better than an full
model fixed effects

Here the null hypothesis
is that the simple model
is the same as the
complex model.

```
B2a <- lme(Richness ~ NAP*fExp, data = RIKZ,  
            random = ~1 | fBeach, method="ML")  
  
B2b <- lme(Richness ~ NAP+fExp, data = RIKZ,  
            random = ~1 | fBeach, method="ML")  
  
anova(B2a, B2b)  
Model df AIC BIC logLik Test L.Ratio p-value  
B2a 1 6 242.1135 252.9535 -115.0567  
B2b 2 5 244.7589 253.7922 -117.3795 1 vs 2 4.645431 0.0311
```

Note the
use of ML!

This is a model comparison problem
This can be done using...
AIC/BIC criteria AND Log-likelihood ratio test

MIXED EFFECTS MODELS

MODEL SELECTION (PART 3B)

```
> B2b <- lme(Richness ~ NAP+fExp, data = RIKZ,  
+ random = ~1 | fBeach, method="ML")  
> B2c <- lme(Richness ~ NAP, data = RIKZ,  
+ random = ~1 | fBeach, method="ML")  
> B2d <- lme(Richness ~ fExp, data = RIKZ,  
+ random = ~1 | fBeach, method="ML")  
> anova(B2b, B2c)  
Model df AIC BIC logLik Test L.Ratio p-value  
B2b 1 5 244.7589 253.7922 -117.3795  
B2c 2 4 249.8291 257.0557 -120.9145 1 vs 2 7.070141 0.0078  
> anova(B2b, B2d)  
Model df AIC BIC logLik Test L.Ratio p-value  
B2b 1 5 244.7589 253.7922 -117.3795  
B2d 2 4 265.4332 272.6599 -128.7166 1 vs 2 22.6743 <.0001
```

NAP AND fExp

NAP

fExp

Should we keep NAP?, what about fBeach?

MIXED EFFECTS MODELS

MODEL SELECTION (PART 4)

Step 4

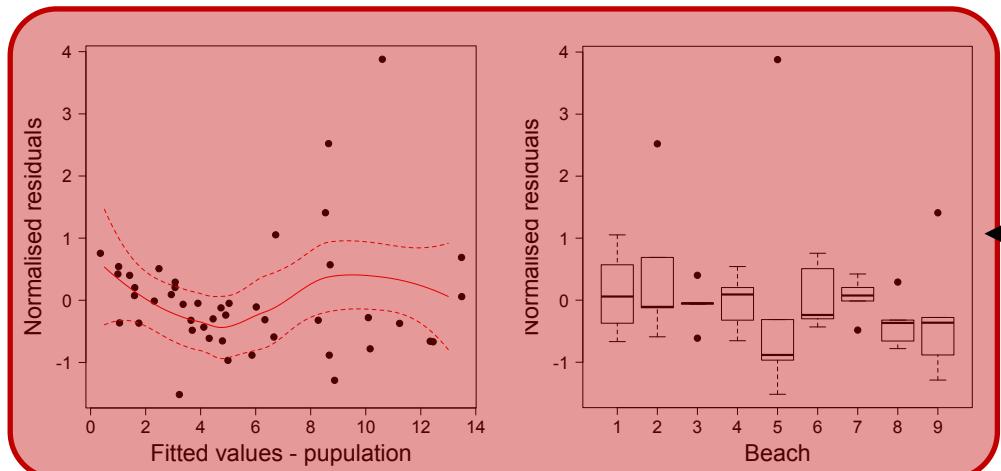
Model Validation

```
> B4 <- update(B2, ~.-NAP:fExp)
> summary(B4)
Linear mixed-effects model fit by REML
  Data: RIKZ
        AIC      BIC      logLik
  240.5538 249.2422 -115.2769
Random effects:
  Formula: ~1 | fBeach
            (Intercept) Residual
  StdDev:    1.907175 3.059089
Fixed effects: Richness ~ NAP + fExp
                  Value Std.Error DF t-value p-value
  (Intercept) 8.601088 1.0594875 35 8.118158 0.0000
  NAP        -2.581708 0.4883901 35 -5.286160 0.0000
  fExp11     -4.532777 1.5755610   7 -2.876929 0.0238
```

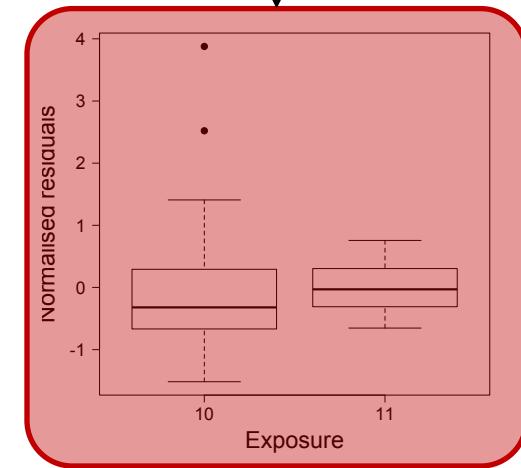
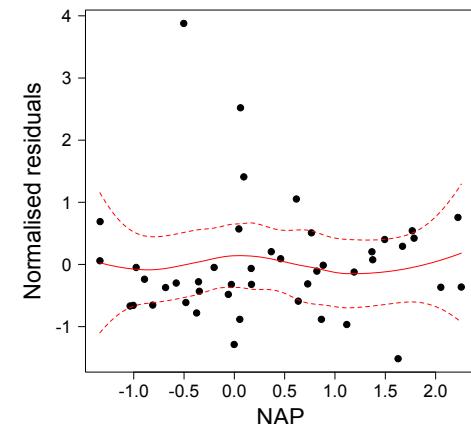
MIXED EFFECTS MODELS

MODEL SELECTION (PART 4)

Is there a structure in the residuals variability?



Heteroscedasticity



So far so good?

Any questions?

Ready to finish?

IN SUMMARY

1. Mixed-effect models that control for pseudo replication can be build by allowing
 - Intercepts to vary - **Random intercept model**
 - Intercepts and the slopes to vary - **Random slope-intercept**

```
lme (Y~Fix.Eff, random=~1+Fix.Eff |  
      Rand.Eff, data=data.In)
```

IN SUMMARY

2. Model selection in mixed models is a two step process:
 - **Part one** – define the best random and/or variance structure
 - Models build using REML
 - **Part two** – define the best fixed effect structure
 - Models build using ML



AARHUS
UNIVERSITY