Mixed Effects Models

Michael Noonan

Biol 520C: Statistical modelling for biological data

Table of contents

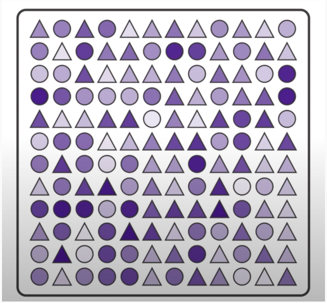
1. Nested Data
2. Mixed Effects Models
3. Fitting Mixed effects models in `R`
4. Effective sample size
5. Pooling and Shrinkage

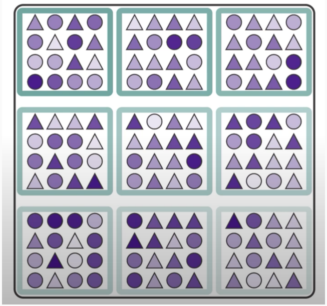
Nested Data

Nested Data

So far we've been fitting models of the general form:
$y_i = \beta_0 + \beta_1 x_i + \ldots + \varepsilon_i$

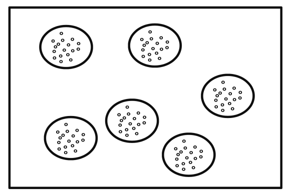
These models assume there is no structure to your data (i.e., obs. come from a single, homogeneous group), but datasets in biology often have some form of overarching structure (e.g., time, space, individual).





Nested Data Cont.

Nested data: Observations belong to nested or hierarchical sub-groups within a population



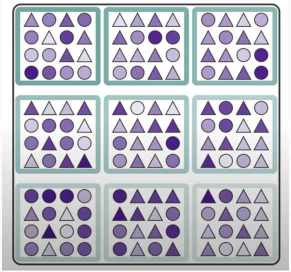
Influence of environmental conditions on plant growth measured across **multiple field sites**

Relationship between movement rates and tick burdens measured across **multiple years**

Influence of fear on foraging duration with multiple levels of 'fear' measured **per individual** (i.e., repeated measures design)

Nested Data Cont.[2]

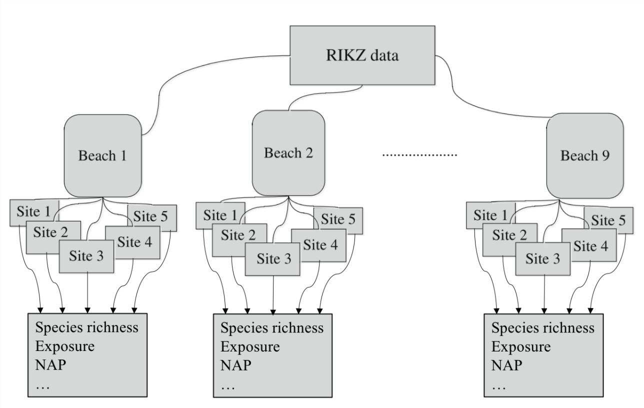
If we suspect that the hierarchical structure of the data influence the outcome of the system, we need to account for this in our model.



The questions is, how?

Modelling nested data

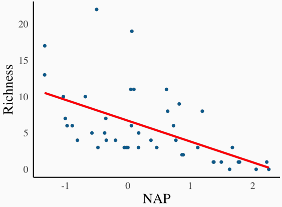
Example: Influence of height of sampling location (NAP) on species richness measured across multiple beaches (Zuur et al. 2007):

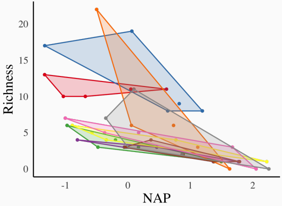


Source: Zuur et al. 2009

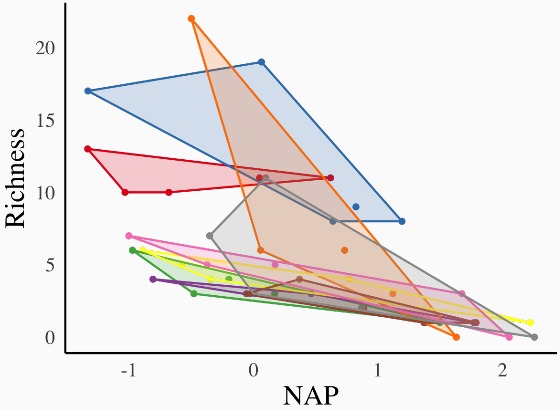
Nested data the wrong way

```
data <- read.delim("RIKZ.txt")

FIT <- lm(Richness ~ NAP, data = data)

summary(FIT)

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

Residuals:
    Min      1Q  Median      3Q     Max
-5.0675 -2.7607 -0.8029  1.3534 13.8723

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.6857     0.6578  10.164 5.25e-13 ***
NAP          -2.8669     0.6307  -4.545 4.42e-05 ***
---

Residual standard error: 4.16 on 43 degrees of freedom
Multiple R-squared:  0.3245,  Adjusted R-squared:  0.3088
F-statistic: 20.66 on 1 and 43 DF,  p-value: 4.418e-05
```





Nested data

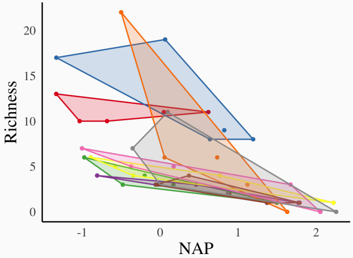


An easy way to account for the nested structure is to add a ‘beach’ term.

$$\text{richness}_i = \beta_0 + \beta_1 \text{NAP}_i + \beta_2 \text{Beach}_i + \varepsilon_i$$

Nested data the wrong way cont.

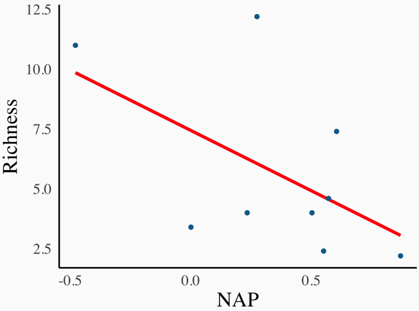
```
data$Beach <- as.factor(data$Beach)

FIT <- lm(Richness ~ NAP + Beach, data = data)

summary(FIT)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.8059     1.3895   7.057 3.22e-08 ***
NAP          -2.4928     0.5023  -4.963 1.79e-05 ***
Beach2        3.0781     1.9720   1.561  0.12755
Beach3       -6.4049     1.9503  -3.284  0.00233 **
Beach4       -6.0329     2.0033  -3.011  0.00480 **
Beach5       -0.8983     2.0105  -0.447  0.65778
Beach6       -5.2231     1.9682  -2.654  0.01189 *
Beach7       -5.4367     2.0506  -2.651  0.01196 *
Beach8       -4.5530     1.9972  -2.280  0.02883 *
Beach9       -3.7820     2.0060  -1.885  0.06770 .
---
Residual standard error: 3.06 on 35 degrees of freedom
Multiple R-squared:  0.7025,  Adjusted R-squared:  0.626
F-statistic: 9.183 on 9 and 35 DF,  p-value: 5.645e-07
```



Our adjusted R^2 is much better... but if we have 9 beaches, and the first site is used as the baseline, this approach requires fitting 8 regression parameters, costing us 8 degrees of freedom! For an effect that might not even be remotely interesting.

Mean regression approach

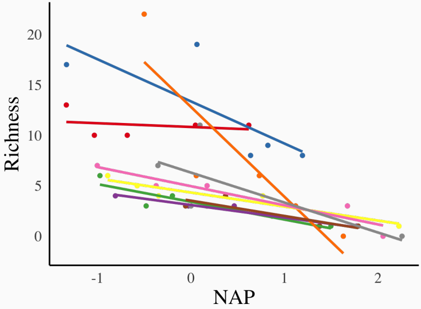
Instead of including a beach term, we could get average values for each beach



... but our analysis only has $n = 9$ and we've lost a lot of the information contained in our original data (plus we've done all that fieldwork for nothing).

Multiple regression approach

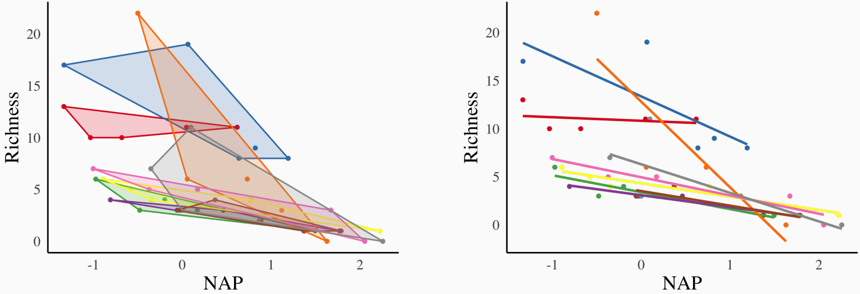
Another approach is to run a separate analysis for each beach



... but each analysis only has $n = 5$, and we have to run multiple tests (risking spurious significance). Clearly, neither of these are great options.

Is there a better way?

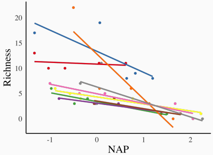
What if we can keep all of our data, and capture the 'beach' effect using a single term?

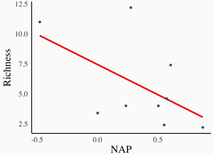


Mixed Effects Models

Mixed Effects Models

The individual regressions may use all of the data, but the results are noisy and hard to interpret. The aggregate data are less noisy, but important differences may have been lost by averaging samples. Mixed effects models can be thought of as lying somewhere in between.





The core of mixed models is that they incorporate fixed and random effects.

Fixed effects are parameters that do not vary.

Random effects are parameters that are themselves random variables.

The Model

In matrix notation a linear mixed effects model can be represented as

$$\boldsymbol{y_i} = X_i\boldsymbol{\beta} + Z_i\boldsymbol{b_i} + \boldsymbol{\varepsilon_i}$$

$\boldsymbol{y_i}$ is the vector of observations ($N \times 1$ vector);

X_i is a matrix of our ‘fixed’ predictor variables ($N \times p$ matrix);

$\boldsymbol{\beta}$ is a vector of fixed effects ($p \times 1$ vector);

Z_i is a matrix of our random predictor variables ($N \times qJ$ matrix for q random effects and J groups);

$\boldsymbol{b_i}$ is a vector of random effects $\sim \mathcal{N}(0, G_i)$ ($qJ \times 1$ vector);

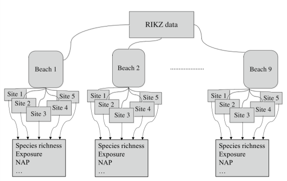
$\boldsymbol{\varepsilon_i}$ is our distribution of errors $\sim \mathcal{N}(0, \sigma_i)$.

$$\overbrace{\mathbf{y}}^{\text{N x 1}} = \overbrace{\underbrace{\mathbf{X}}_{\text{N x p}} \quad \underbrace{\boldsymbol{\beta}}_{\text{p x 1}}}^{\text{N x 1}} + \overbrace{\underbrace{\mathbf{Z}}_{\text{N x qJ}} \quad \underbrace{\boldsymbol{b}}_{\text{qJ x 1}}}^{\text{N x 1}} + \overbrace{\varepsilon}^{\text{N x 1}}$$

The Model Cont.

$$\overbrace{\mathbf{y}}^{\text{N x 1}} = \overbrace{\underbrace{\mathbf{X}}_{\text{N x p}}\quad\underbrace{\boldsymbol{\beta}}_{\text{p x 1}}}^{\text{N x 1}} + \overbrace{\underbrace{\mathbf{Z}}_{\text{N x qJ}}\quad\underbrace{\boldsymbol{b}}_{\text{qJ x 1}}}^{\text{N x 1}} + \overbrace{\varepsilon}^{\text{N x 1}}$$

In these data richness was measured at 5 sites on 9 beaches ($N = 45$). We're modelling one predictor variable (NAP) and a fixed intercept ($p = 2$). Richness was measured at 9 beaches ($J = 9$), and there's 1 random intercept ($q = 1$).



$$\overbrace{\mathrm{y}}^{\text{45 x 1}} = \overbrace{\underbrace{\mathbf{X}}_{\text{45 x 2}}\quad\underbrace{\boldsymbol{\beta}}_{\text{2 x 1}}}^{\text{45 x 1}} + \overbrace{\underbrace{\mathbf{Z}}_{\text{45 x 9}}\quad\underbrace{\boldsymbol{b}}_{\text{9 x 1}}}^{\text{45 x 1}} + \overbrace{\varepsilon}^{\text{45 x 1}}$$

The Model Cont.

$$\overbrace{\mathrm{y}}^{45 \text{ x } 1} = \overbrace{\underbrace{\mathbf{X}}_{45 \text{ x } 2} \quad \underbrace{\boldsymbol{\beta}}_{2 \text{ x } 1}}^{45 \text{ x } 1} + \overbrace{\underbrace{\mathbf{Z}}_{45 \text{ x } 9} \quad \underbrace{\boldsymbol{b}}_{9 \text{ x } 1}}^{45 \text{ x } 1} + \overbrace{\varepsilon}^{45 \text{ x } 1}$$

$$\mathbf{y} = \begin{bmatrix} \text{Rich.} \\ 11 \\ 10 \\ \dots \\ 2 \end{bmatrix} \begin{matrix} n_{ij} \\ 1 \\ 2 \\ \dots \\ 45 \end{matrix} \quad \mathbf{X} = \begin{bmatrix} \text{Int.} & \text{NAP} \\ 1 & 0.045 \\ 1 & -1.036 \\ \dots & \dots \\ 1 & 0.865 \end{bmatrix} \boldsymbol{\beta} = \begin{bmatrix} \beta_{Intercept} \\ \beta_{NAP} \end{bmatrix}$$

Z is a matrix of 0s and 1s telling us which beach the richness data are from (each row is a richness record, each column is a beach).

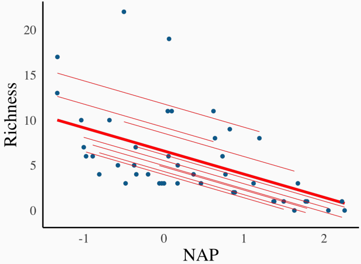
b is a column vector similar to $\boldsymbol{\beta}, \boldsymbol{b} \sim \mathcal{N}(\mathbf{0}, \mathbf{G})$, where G is the variance-covariance matrix of the random effects.

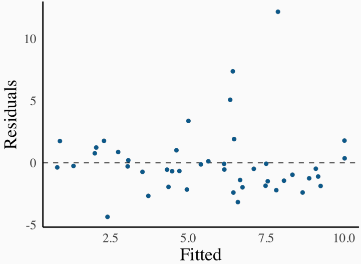
Fitting Mixed effects models in R

Random Intercept Model in R

THE UNIVERSITY OF BRITISH COLUMBIA
Okanagan Campus

```
library(nlme)

FIT <- lme(Richness ~ NAP,
random = ~1 | Beach, data = data)

summary(FIT)

Linear mixed-effects model fit by REML
 Data: data
        AIC      BIC    logLik
  247.4802 254.525 -119.7401

Random effects:
 Formula: ~1 | Beach
        (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

Standardized Within-Group Residuals:
       Min         Q1        Med          Q3
            Max
-1.4227495 -0.4848006 -0.1576462  0.2518966
       3.9793918

Number of Observations: 45
Number of Groups: 9
```





Random Intercept Model in R

```
library(nlme)

FIT <- lme(Richness ~ NAP,
random = ~1 | Beach, data = data)

summary(FIT)

Linear mixed-effects model fit by REML
 Data: data
       AIC      BIC    logLik
  247.4802 254.525 -119.7401

Random effects:
 Formula: ~1 | Beach
        (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

Standardized Within-Group Residuals:
       Min         Q1        Med         Q3
            Max
-1.4227495 -0.4848006 -0.1576462  0.2518966
      3.9793918

Number of Observations: 45
Number of Groups: 9
```

$$\overbrace{\mathrm{y}}^{45 \times 1} = \overbrace{\underbrace{\mathbf{X}}_{45 \times 2} \underbrace{\boldsymbol{\beta}}_{2 \times 1}}^{45 \times 1} + \overbrace{\underbrace{\mathbf{Z}}_{45 \times 9} \underbrace{\boldsymbol{b}}_{9 \times 1}}^{45 \times 1} + \overbrace{\boldsymbol{\varepsilon}}^{45 \times 1}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{Intercept} \\ \beta_{NAP} \end{bmatrix} \approx \begin{bmatrix} 6.58 \\ -2.57 \end{bmatrix}$$

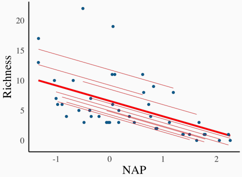
$\boldsymbol{b} \sim \mathcal{N}(\mathbf{0}, \mathbf{G})$, where $\mathrm{G} \approx 2.94^2 \approx 8.67$

```
 FIT$coefficients$random$Beach

  (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
```

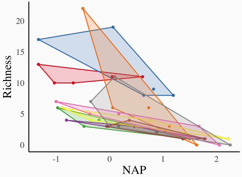
```
sd(FIT$coefficients$random$Beach)
[1] 2.664828
```

Random Intercept and slope model

The random intercept model we just fit allows for correlation in Richness *within* beaches, but assumes the relationship between NAP and Richness is *fixed across* beaches.



What if the relationship between NAP and Richness differs across beaches?

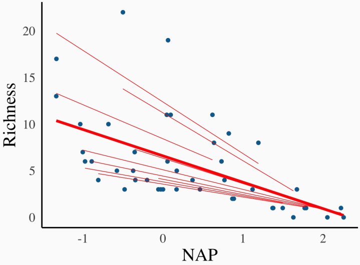


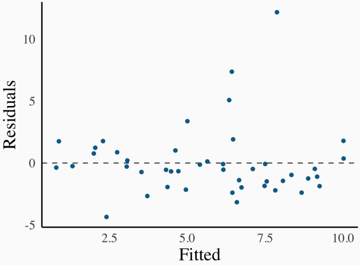
From a fixed effects perspective, this implies including an interaction term between NAP and Beach (but at the cost of needing to fit 17 additional parameters!).

In fact, we can apply the same principle of including a random intercept to the slope.

Random Intercept and slope model in R

```
FIT <- lme(Richness ~ NAP,
random = ~ 1+ NAP | Beach, data = data)

summary(FIT)

Linear mixed-effects model fit by REML
 Data: data
       AIC      BIC    logLik
  244.3839 254.9511 -116.1919

Random effects:
 Formula: ~NAP | Beach
 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
 Correlation:
    (Intr)
NAP -0.819
```





Random Intercept and slope model in R

```
FIT <- lme(Richness ~ NAP,
random = ~ 1+ NAP | Beach, data = data)

summary(FIT)

Linear mixed-effects model fit by REML
 Data: data
       AIC      BIC    logLik
  244.3839 254.9511 -116.1919

Random effects:
 Formula: ~NAP | Beach
 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
 Correlation:
    (Intr)
NAP -0.819
```

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{Intercept} \\ \beta_{NAP} \end{bmatrix} \approx \begin{bmatrix} 6.59 \\ -2.83 \end{bmatrix}$$

$$\boldsymbol{b} \sim \mathcal{N}(\mathbf{0}, \mathbf{G}), \text{ where } \mathbf{G} = \begin{bmatrix} \sigma^2_{int} & \sigma^2_{int,slope} \\ \sigma^2_{int,slope} & \sigma^2_{slope} \end{bmatrix} =$$

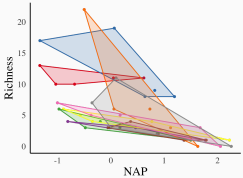
$$\begin{bmatrix} 3.54^2 & -0.99 * 3.54 * 1.71 \\ -0.99 * 3.54 * 1.71 & 1.71^2 \end{bmatrix}$$

```
getVarCov(FIT)
Random effects variance covariance matrix
            (Intercept)     NAP
(Intercept)     12.5960 -6.0268
NAP             -6.0268  2.9411
  Standard Deviations: 3.5491 1.715


 FIT$coefficients$random$Beach

  (Intercept)        NAP
1   1.8323503 -0.8262398
2   5.7747909 -2.7067858
3  -2.7820638  1.3243120
4  -3.0262848  1.4440673
5   4.6114440 -2.3072951
6  -2.1624275  1.0543530
7  -2.5057621  1.1856990
8  -1.4888117  0.7231759
9  -0.2532354  0.1087136
```

Effective sample size

Non-independence within 'blocks'

The reason we started including random effects was to control for non-independence of measures from the same beach.



This implies that there is some measurable correlation between samples from the same beach.

When data are correlated, each new datapoint does not bring a full independent datapoint worth of information with it, and $N \neq N_{\text{effective}}$ (we'll cover this in detail in a later lecture).

Induced correlation

Defining a random effect structure to the model induces a correlation structure defined as: $\frac{g^2}{g^2+\sigma^2}$ (also called interclass correlation, ρ)

1. ρ can help you determine whether a linear mixed model is even necessary (e.g., if ρ is zero, observations within clusters are no more similar than observations from different clusters and random effects aren't necessary).
2. It can be theoretically meaningful to understand how much of the overall variation in the response is explained simply by clustering.

```
FIT <- lme(Richness ~ NAP,
random = ~1 | Beach, data = data)

Random effects:
 Formula: ~1 | Beach
        (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
```

$$\rho = \frac{g^2}{g^2+\sigma^2}$$

$$g = 2.94 \ \& \ \sigma = 3.06$$

$$= \frac{2.94^2}{2.94^2+3.06^2} = 0.48$$

Interclass correlation and $N_{\text{effective}}$

An interclass correlation of 0.48 tells us that samples from the same beach are ∼50% related to one another (or $\sim 1/2$ of the information in each new data point is duplicate).

In other words, the random effect structure was clearly needed.

This means our sample size needs to be adjusted for this 'design effect':

$$\text{design effect} = 1 + (n-1)\rho = 1 + (5-1) \times 0.48 = 2.92$$

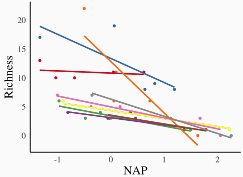
$$N_{\text{effective}} = \frac{N \times n}{\text{design effect}} = \frac{9 \times 5}{2.92} = 15.41$$

15.41 is higher than averaging samples across beaches ($n = 9$), or running 9 regressions on individual beaches ($n = 5$), but lower than the full sample size of 45.

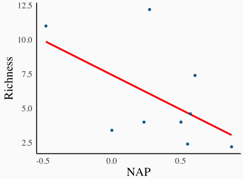
Pooling and Shrinkage

Partial pooling

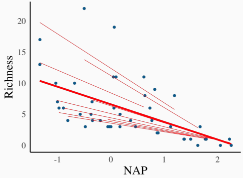
One of the options for controlling for non-independence was to run a separate model for each beach, meaning the parameters for each beach are estimated independently (i.e., **no pooling**).



Another option was to average species richness within beaches and fit the model to the mean data (i.e., **complete pooling**).

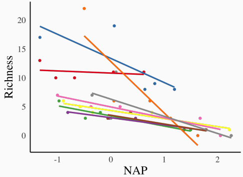


Fitting a linear mixed model to all the data, using beach as a random factor, allows for different intercepts and slopes for each beach, but unlike the no-pooling analysis, these estimates combine information from the other beaches (i.e., **partial pooling**).

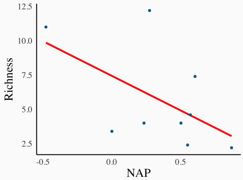


Partial pooling cont.

The consequence of partial pooling in a linear mixed model is that group specific intercepts are pulled toward the mean intercept and the group specific slopes are pulled toward the mean slope.



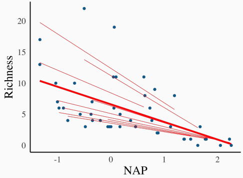
The result is that the *differences* in parameter estimates among sites are shrunk toward zero (called ‘shrinkage’).



A consequence of this shrinkage is that the variance of the intercept estimates or of the slope estimates is smaller than that in the no-pooling analysis.

The benefit is that groups with very little data can lean on the information from other groups.

The downside is that groups with very little data lean on the information from other groups.



References

Zuur, A et al. (2009). Mixed effects models and extensions in ecology with R. Springer.