# Biostatistics-Lecture 15
# Linear Mixed Model

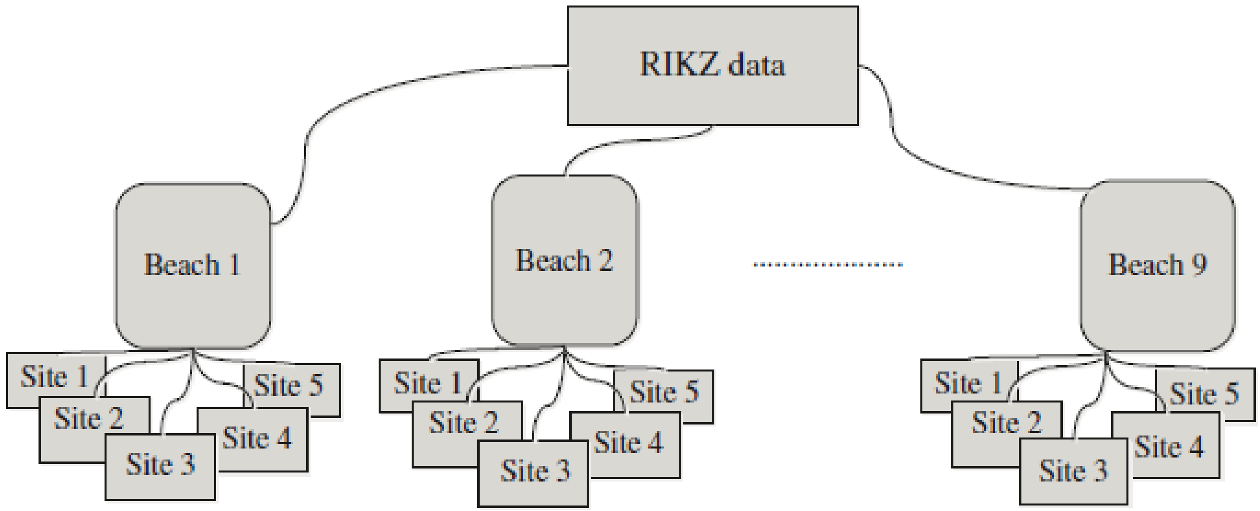
Ruibin Xi
Peking University
School of Mathematical Sciences

# An example

- Zuur et al. (2007) measured marine benthic data
  - 9 inter-tidal area were measured
  - 5 samples were taken in each area
- Question: whether species richness is related to
  - NAP: the height of the sampling station compared to the mean tidal level
  - Exposure: an index composed of wave action, length of the surf zone, slope, grain size, and the depth of the anaerobic layer.

# An example

- Zuur et al. (2007) measured marine benthic data
  - 9 inter-tidal area were measured
  -
- Q ...ed to
  - ...red to
  - ...ength ...of the anaerobic layer.



# An example

- Zuur et al. (2007) measured marine benthic data
  - 9
  - 5
- Que … lated to
  - N … npared to th
  - E … n, length o … epth of the anaerobic layer.

| | Sample | Richness | Exposure | NAP | Beach |
|---|---|---|---|---|---|
| 1 | 1 | 11 | 10 | 0.045 | 1 |
| 2 | 2 | 10 | 10 | -1.036 | 1 |
| 3 | 3 | 13 | 10 | -1.336 | 1 |
| 4 | 4 | 11 | 10 | 0.616 | 1 |
| 5 | 5 | 10 | 10 | -0.684 | 1 |
| 6 | 6 | 8 | 8 | 1.190 | 2 |
| 7 | 7 | 9 | 8 | 0.820 | 2 |
| 8 | 8 | 8 | 8 | 0.635 | 2 |
| 9 | 9 | 19 | 8 | 0.061 | 2 |
| 10 | 10 | 17 | 8 | -1.334 | 2 |

Exposure can take 3 possible values, 8, 10, and 11.
There are 5, 20 and 20 samples taking values 8, 10, and 11, respectively

# A first model

- We may first try the linear model

$$R_{ij} = \alpha + \beta_1 \times NAP_{ij} + \beta_2 \times Exposure_i + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim N(0, \sigma^2)$$

$R_{ij}$ is the species richness at site $j$ on beach $i$.

$NAP_{ij}$ the corresponding NAP value

$Exposure_i$ the exposure on beach $i$,

$\varepsilon_{ij}$ the unexplained information.

# 2-stage Analysis method

- 1st stage: model species richness and NAP for each beach

$$R_{ij} = \alpha + \beta_i \times NAP_{ij} + \varepsilon_{ij} \qquad j = 1, \ldots, 5$$

$$\begin{pmatrix} R_{i1} \\ R_{i2} \\ R_{i3} \\ R_{i4} \\ R_{i5} \end{pmatrix} = \begin{pmatrix} 1 & NAP_{i1} \\ 1 & NAP_{i1} \\ 1 & NAP_{i1} \\ 1 & NAP_{i1} \\ 1 & NAP_{i1} \end{pmatrix} \times \begin{pmatrix} \alpha \\ \beta_i \end{pmatrix} + \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \\ \varepsilon_{i4} \\ \varepsilon_{i5} \end{pmatrix} \Leftrightarrow \mathbf{R}_i = \mathbf{Z}_i \times \boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i$$

- We can get beta as

```
-0.37 -4.17 -1.75 -1.24 -8.90 -1.38 -1.51 -1.89 -2.96
```

# 2-stage analysis method

- $2^{nd}$ stage: the estimated regression coefficient are modeled as a function of exposure

$$\hat{\beta}_i = \eta + \tau \times Exposure_i + b_i \qquad i = 1, \ldots, 9$$

$$\begin{pmatrix} -0.37 \\ -4.17 \\ -1.75 \\ -1.24 \\ -8.90 \\ -1.38 \\ -1.51 \\ -1.89 \\ -2.96 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \end{pmatrix} \times \begin{pmatrix} \eta \\ \tau \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \\ b_6 \\ b_7 \\ b_8 \\ b_9 \end{pmatrix} \Leftrightarrow \hat{\boldsymbol{\beta}}_i = \mathbf{K}_i \times \boldsymbol{\gamma} + \mathbf{b}_i \qquad i = 1, \ldots, 9$$

Code 8 and 10 as 0, 11 as 1

# 2-stage analysis method

```
Call:
lm(formula = Beta ~ factor(ExposureBeach), data = RIKZ)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.2386 -0.2778  0.0890  0.6940  3.2897 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)  
(Intercept)              -3.662      1.099  -3.332   0.0126 *
factor(ExposureBeach)b    2.184      1.649   1.325   0.2268  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 2.458 on 7 degrees of freedom
Multiple R-squared:  0.2005,    Adjusted R-squared:  0.08625 
F-statistic: 1.755 on 1 and 7 DF,  p-value: 0.2268
```

Exposure is not significant

# 2-stage analysis method

- Disadvantage:
  - Summarize all data from a beach with one parameter
  - In the 2nd step, we are analyzing the regression parameters, not the observed data (not modeling the variable of interest)
  - The number of observations used to calculate the summary statistic is not used in the 2nd step.

# Linear Mixed effect model

- The model

$$\mathbf{R}_i = \mathbf{X}_i \times \beta + \mathbf{Z}_i \times \mathbf{b}_i + \varepsilon_i$$

$$\begin{aligned}
&\mathbf{Y}_i = \mathbf{X}_i \times \beta + \mathbf{Z}_i \times \mathbf{b}_i + \varepsilon_i \\
&\mathbf{b}_i \sim \mathrm{N}(\mathbf{0}, \mathbf{D}) \\
&\varepsilon_i \sim \mathrm{N}(\mathbf{0}, \mathbf{\Sigma}_i) \\
&\mathbf{b}_1, \cdots \mathbf{b}_\mathrm{N}, \varepsilon_1, \cdots, \varepsilon_\mathrm{N} \text{ independent}
\end{aligned}$$

Fixed effect term

Random effect term

# Linear Mixed effect model



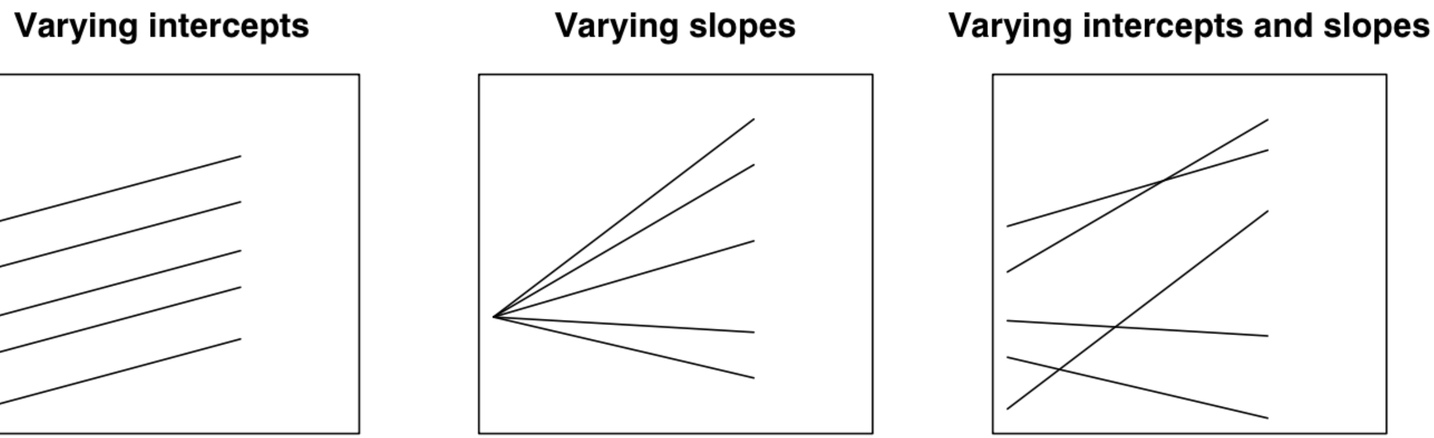



Figure 11.1 *Linear regression models with (a) varying intercepts* ($y = \alpha_j + \beta x$)*, (b) varying slopes* ($y = \alpha + \beta_j x$)*, and (c) both* ($y = \alpha_j + \beta_j x$)*. The varying intercepts correspond to group indicators as regression predictors, and the varying slopes represent interactions between* $x$ *and the group indicators.*

# The random intercept model

- Model the richness as a linear function of NAP
  - Intercept is allowed to change per beach

$$R_{ij} = \alpha + \beta_1 \times Beach_i + \beta_2 \times NAP_{ij} + \varepsilon_{ij}$$

$$\begin{pmatrix} R_{i1} \\ R_{i2} \\ R_{i3} \\ R_{i4} \\ R_{i5} \end{pmatrix} = \begin{pmatrix} 1 & NAP_{i1} \\ 1 & NAP_{i2} \\ 1 & NAP_{i3} \\ 1 & NAP_{i4} \\ 1 & NAP_{i5} \end{pmatrix} \times \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \times b_i + \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \\ \varepsilon_{i4} \\ \varepsilon_{i5} \end{pmatrix} \Leftrightarrow \mathbf{R}_i = \mathbf{X}_i \times \boldsymbol{\beta} + \mathbf{Z}_i \times \mathbf{b}_i$$

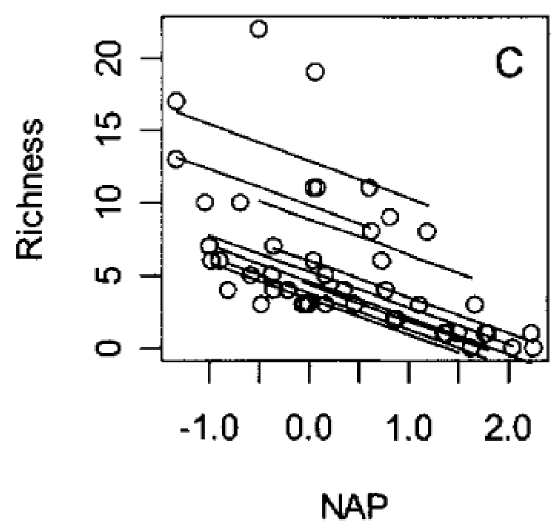
$b_i$ are normally distributed: $N(0, d^2)$.

$$\Sigma_i = \sigma^2 I_5$$

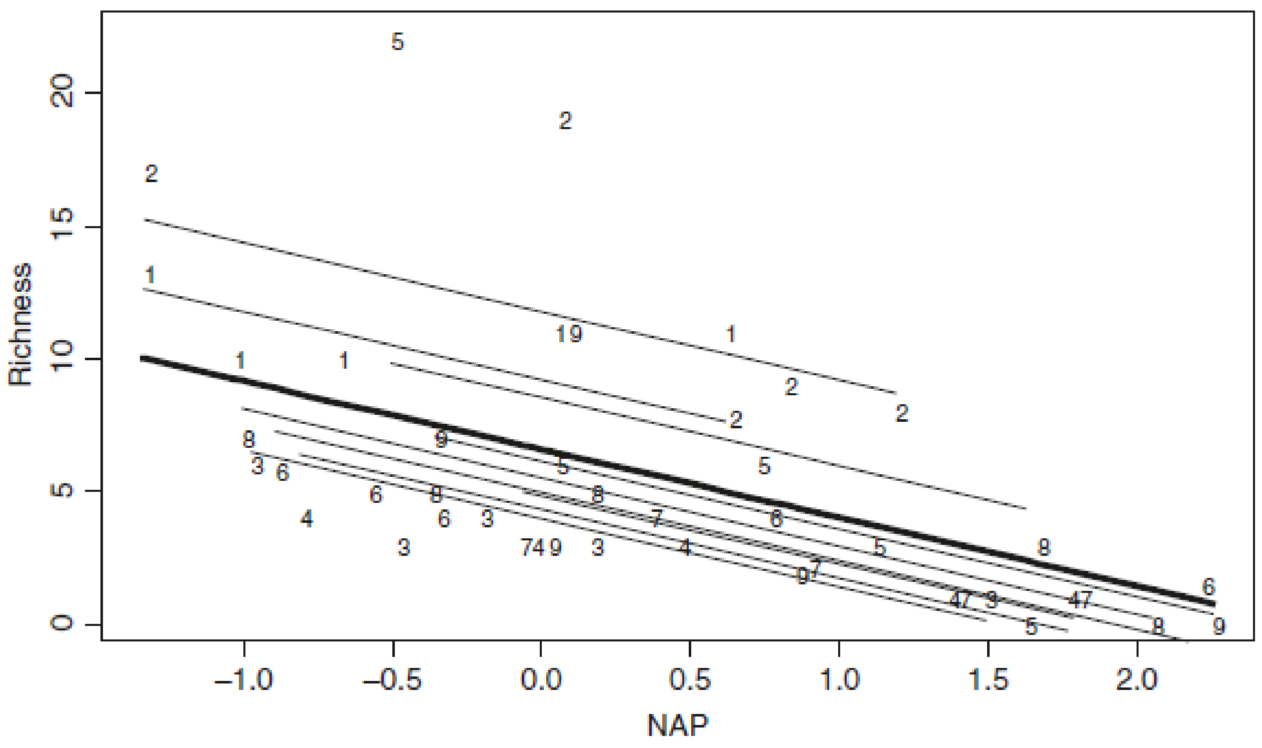
# The random intercept model

- $y_{ij} = \alpha + \beta x_{ij} + a_j + \varepsilon_{ij}$ where $a_j \sim N(0, \sigma_a^2)$ and $\varepsilon_{ij} \sim N(0, \sigma^2)$
- model2 <- lme (richness ~ NAP, random = ~1|beach, method = “REML”, data)

i.e. a model that fits the same slope for each level of the random factor (fitted by REML by default)



# The random intercept model



# The random intercept and slope model

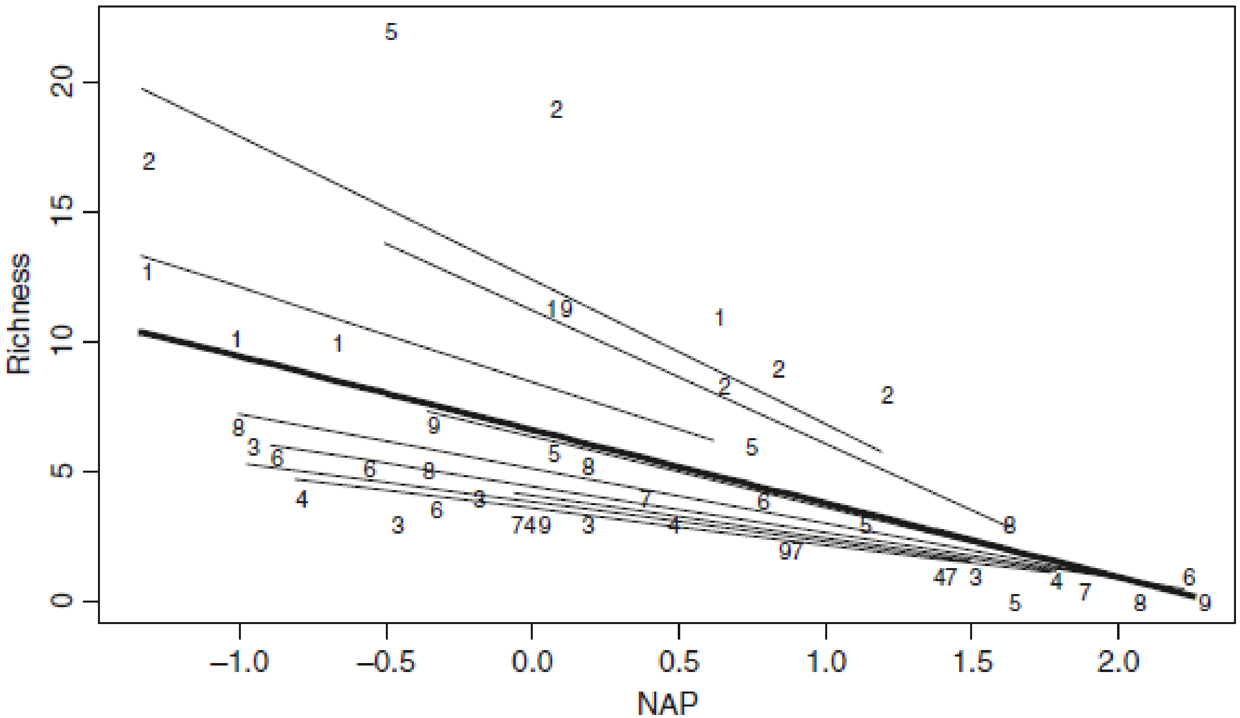
- The model

$$\begin{pmatrix} R_{i1} \\ R_{i2} \\ R_{i3} \\ R_{i4} \\ R_{i5} \end{pmatrix} = \begin{pmatrix} 1 & NAP_{i1} \\ 1 & NAP_{i2} \\ 1 & NAP_{i3} \\ 1 & NAP_{i4} \\ 1 & NAP_{i5} \end{pmatrix} \times \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} 1 & NAP_{i1} \\ 1 & NAP_{i2} \\ 1 & NAP_{i3} \\ 1 & NAP_{i4} \\ 1 & NAP_{i5} \end{pmatrix} \times \begin{pmatrix} b_{i1} \\ b_{i2} \end{pmatrix} + \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \\ \varepsilon_{i4} \\ \varepsilon_{i5} \end{pmatrix}$$

$$\begin{pmatrix} b_{i1} \\ b_{i2} \end{pmatrix} \sim N(\mathbf{0}, \mathbf{D})$$

$$\mathbf{D} = \begin{pmatrix} d_{11}^2 & d_{12} \\ d_{12} & d_{22}^2 \end{pmatrix}$$

# The random intercept and slope model



# Induced correlation

- The model

$$\mathbf{Y}_i = \mathbf{X}_i \times \boldsymbol{\beta} + \mathbf{Z}_i \times \mathbf{b}_i + \boldsymbol{\varepsilon}_i$$

- Marginal distribution of Y is

$$\mathbf{Y}_i \sim N(\mathbf{X}_i \times \boldsymbol{\beta}, \mathbf{V}_i)$$

$$\mathbf{V}_i = \mathbf{Z}_i \times \mathbf{D} \times \mathbf{Z}_i' + \boldsymbol{\Sigma}_i$$

- In case of the random intercept model, we have

# Induced correlation

- The model
- Margi

$$
\mathbf{V}_i = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \times d^2 \times \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \end{pmatrix} + \sigma^2 \times \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & & & \vdots \\ \vdots & & 1 & & \vdots \\ \vdots & & & 1 & 0 \\ 0 & \cdots & \cdots & 0 & 1 \end{pmatrix}
$$

$$
= \begin{pmatrix} \sigma^2 + d^2 & d^2 & d^2 & d^2 & d^2 \\ d^2 & \sigma^2 + d^2 & d^2 & d^2 & d^2 \\ d^2 & d^2 & \sigma^2 + d^2 & d^2 & d^2 \\ d^2 & d^2 & d^2 & \sigma^2 + d^2 & d^2 \\ d^2 & d^2 & d^2 & d^2 & \sigma^2 + d^2 \end{pmatrix}
$$

- In cas … el, we have

# Induced correlation

- For the random intercept and slope model, we have

$$\mathbf{V}_i = \begin{pmatrix} 1 & NAP_{i1} \\ 1 & NAP_{i2} \\ 1 & NAP_{i3} \\ 1 & NAP_{i4} \\ 1 & NAP_{i5} \end{pmatrix} \times \begin{pmatrix} d_{11}^2 & d_{21} \\ d_{12} & d_{22}^2 \end{pmatrix} \times \begin{pmatrix} 1 & NAP_{i1} \\ 1 & NAP_{i2} \\ 1 & NAP_{i3} \\ 1 & NAP_{i4} \\ 1 & NAP_{i5} \end{pmatrix} + \sigma^2 \times \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & & & \vdots \\ \vdots & & 1 & & \vdots \\ \vdots & & & 1 & 0 \\ 0 & \cdots & \cdots & 0 & 1 \end{pmatrix}$$

$$\text{var}(Y_{ij}) = d_{11}^2 + 2 \times NAP_{ij} \times d_{12} + NAP_{ij}^2 \times d_{22}^2 + \sigma^2$$

$$\text{cov}(Y_{ij}, Y_{ik}) = d_{11}^2 + (NAP_{ij} + NAP_{ik}) \times d_{12} + NAP_{ij} \times NAP_{ik} \times d_{22}^2$$

# The marginal model

- If we integrate out b, we get

$$\mathbf{Y}_i \sim N(\mathbf{X}_i \times \beta, \mathbf{V}_i)$$

$$\mathbf{V}_i = \mathbf{Z}_i \times \mathbf{D} \times \mathbf{Z}_i' + \mathbf{\Sigma}_i$$

- In general, if we don't have the random effects,

$$\mathbf{V}_i = \mathbf{\Sigma}_i = \begin{pmatrix} \sigma^2 & c_{21} & c_{31} & c_{41} & c_{51} \\ c_{21} & \sigma^2 & c_{32} & c_{42} & c_{52} \\ c_{31} & c_{32} & \sigma^2 & c_{43} & c_{53} \\ c_{41} & c_{42} & c_{43} & \sigma^2 & c_{54} \\ c_{54} & c_{52} & c_{53} & c_{54} & \sigma^2 \end{pmatrix}$$

General correlation matrix

# The marginal model

- If we integrate out b, we get

$$\mathbf{Y}_i \sim N(\mathbf{X}_i \times \beta, \mathbf{V}_i)$$

$$\mathbf{V}_i = \mathbf{Z}_i \times \mathbf{D} \times \mathbf{Z}_i' + \mathbf{\Sigma}_i$$

- In general, if we don't have the random effects,

$$\mathbf{V}_i = \Sigma_i = \begin{pmatrix} \sigma^2 & \varphi & \varphi & \varphi & \varphi \\ \varphi & \sigma^2 & \varphi & \varphi & \varphi \\ \varphi & \varphi & \sigma^2 & \varphi & \varphi \\ \varphi & \varphi & \varphi & \sigma^2 & \varphi \\ \varphi & \varphi & \varphi & \varphi & \sigma^2 \end{pmatrix}$$

Compound symmetric structure

# REML

- REML: restricted maximum likelihood estimate
- In the simple linear regression

$$Y_i = \alpha + \beta \times X_i + \varepsilon_i$$

- An unbiased estimate of the variance is

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} \times X_i)^2$$

- But the MLE is

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} \times X_i)^2$$

Biased!!

# REML

- REML works as follows

$$Y_i = \alpha + \beta \times X_i + \varepsilon_i$$

- Written in matrix form

$$Y_i = \mathbf{X}_i \times \boldsymbol{\beta} + \varepsilon_i$$

- The normality assumption implies

$$Y_i \sim N(\mathbf{X}_i \times \boldsymbol{\beta}, \sigma^2)$$

- Take A of dimension n×(n-2), such that A’ and X are orthogonal,

$$\mathbf{A}' \times \mathbf{Y} = \breve{\mathbf{A}}' \times \mathbf{X} \times \boldsymbol{\beta} + \mathbf{A}' \times \boldsymbol{\varepsilon}$$

$$\mathbf{A}' \times \mathbf{Y} \sim N(\mathbf{0}, \sigma^2 \times \mathbf{A}' \times \mathbf{A})$$

Maximize this we get REML

# REML

- For the linear mixed model

$$\mathbf{Y}_i \sim N(\mathbf{X}_i \times \boldsymbol{\beta}, \mathbf{V}_i) \qquad \mathbf{V}_i = \mathbf{Z}_i \times \mathbf{D} \times \mathbf{Z}_i' + \boldsymbol{\Sigma}_i$$

$$\ln L(\mathbf{Y}_i, \mathbf{X}_i, \boldsymbol{\theta}) = -\frac{n}{2}\ln 2\pi - \frac{1}{2}\sum_{i=1}^{n}\ln|\mathbf{V}_i| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{Y}_i - \mathbf{X}_i \times \boldsymbol{\beta})' \times \mathbf{V}_i^{-1} \times (\mathbf{Y}_i - \mathbf{X}_i \times \boldsymbol{\beta})$$

- Putting all observations together

$$\mathbf{Y} \sim \mathrm{N}(\mathbf{X} \times \boldsymbol{\beta}, \mathbf{V})$$

- Similarly, we can take A with $\mathbf{A}' \times \mathbf{X} = \mathbf{0}$
- Thus $\mathbf{A}' \times \mathbf{Y} \sim \mathrm{N}(\mathbf{0}, \mathbf{A}' \times \mathbf{V} \times \mathbf{A})$

# REML

| Mixed model with NAP as fixed covariate and random intercept | | |
| --- | --- | --- |
| Parameter | Estimate using ML | Estimate using REML |
| Fixed intercept | 6.58 (1.05) | 6.58 (1.09) |
| Fixed slope NAP | –2.57 (0.49) | –2.56 (0.49) |
| Variance random intercept | 7.50 | 8.66 |
| Residual variance | 9.11 | 9.36 |
| AIC | 249.82 | 247.48 |
| BIC | 257.05 | 254.52 |
| Mixed model with NAP and exposure as fixed covariate and random intercept | | |
| Fixed intercept | 8.60 (0.96) | 8.60 (1.05) |
| Fixed slope NAP | –2.60 (0.49) | –2.58 (0.48) |
| Fixed Exposure level | –4.53 (1.43) | –4.53 (1.57) |
| Variance random intercept | 2.41 | 3.63 |
| Residual variance | 9.11 | 9.35 |
| AIC | 244.75 | 240.55 |
| BIC | 253.79 | 249.24 |