02409 Multivariate Statistics

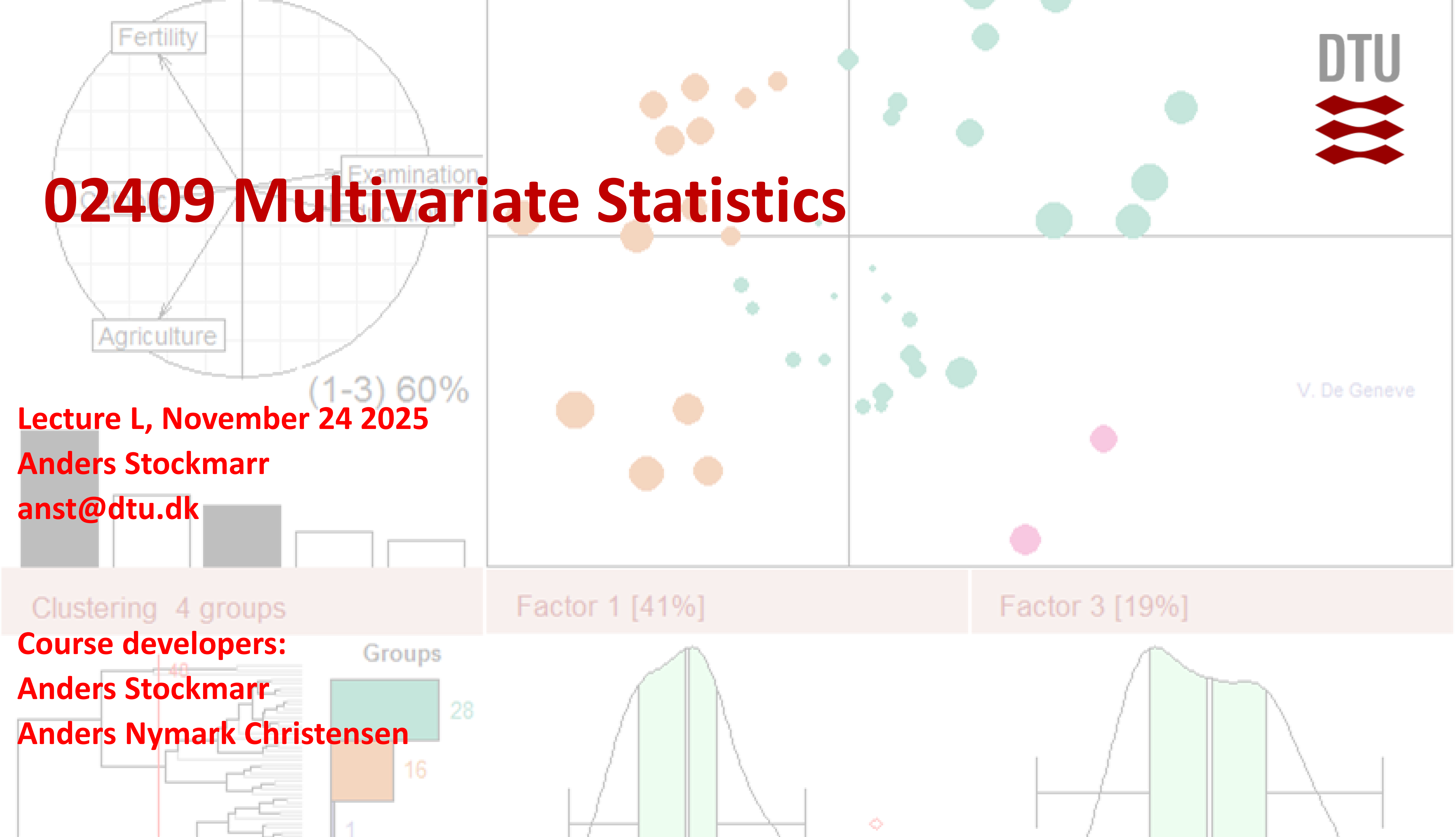
Lecture L, November 24 2025

Anders Stockmarr
anst@dtu.dk

Course developers:

Anders Stockmarr

Anders Nymark Christensen



Agenda

- Exam
- Topics for the last lecture
- Course evaluation
- Multivariate ANOVA
- Bartlett’s test
- Covariance hypotheses:

  Repeated Measurements Models

Exam

- 4 hours multiple choice
  - On DTU
  - Only digital
  - All aids
    - including internet
  - We will use the Digital Exam platform
- Contents
  - There will be questions that will require the use of statistical software – **R**.
  - At the exercises, you can get a recap of how to load data into **R**.
- Packages used at the exercises – best install them before the exam:

```
psych, geigen, CCP, readr, MASS, Rfast, car
```

- `MASS` is pre-installed in R, and only needs to be activated.

Last lecture

- Topics from the entire curriculum
- More detailed in areas where students have requested it;
  - *send me an email as soon as possible.*
- *Last years exam*

Course evaluation

- Deadline: **28th november 2023 23:59**
- Great help for developing the course
  - The course is under reconstruction;
  - E.g. ideas for feedback during the course;
- CCC
  - (Caring)
  - Constructive
  - Concrete

The Multivariate General Linear Model

$\boldsymbol{Y}_1 \cdots \boldsymbol{Y}_n$ independent random observations

$$\boldsymbol{Y}_i \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}), \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \cdots & \sigma_{1p} \\ \vdots & & \vdots \\ \sigma_{p1} & \cdots & \sigma_p^2 \end{bmatrix}$$

$$\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{Y}_1^T \\ \vdots \\ \boldsymbol{Y}_n^T \end{bmatrix} = \begin{bmatrix} Y_{11} & \cdots & Y_{1p} \\ \vdots & & \vdots \\ Y_{n1} & \cdots & Y_{np} \end{bmatrix} = [\boldsymbol{Y}_{,1} \quad \cdots \quad \boldsymbol{Y}_{,p}]$$

$$E(\boldsymbol{Y}) = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \theta_{11} & \cdots & \theta_{1p} \\ \vdots & & \vdots \\ \theta_{k1} & \cdots & \theta_{kp} \end{bmatrix} = \boldsymbol{x\theta} = \begin{bmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_n \end{bmatrix} [\boldsymbol{\theta}_{,1} \quad \cdots \quad \boldsymbol{\theta}_{,p}]$$

$E\left(\boldsymbol{Y}_i^T\right) = \boldsymbol{x}_i\boldsymbol{\theta}$ with $\boldsymbol{x}_i = [x_{i1} \quad \cdots \quad x_{ik}]$

$$V(\boldsymbol{Y}) = \boldsymbol{I}_n \otimes \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma} & & \\ & \ddots & \\ & & \boldsymbol{\Sigma} \end{bmatrix}$$

$\hat{\boldsymbol{\theta}} = (\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{Y}$; ***calculated one column at a time***: $\hat{\boldsymbol{\theta}}_{,\boldsymbol{j}} = (\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{Y}_{,j}$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n-k}\left(\boldsymbol{Y} - \boldsymbol{x}\hat{\boldsymbol{\theta}}\right)^T\left(\boldsymbol{Y} - \boldsymbol{x}\hat{\boldsymbol{\theta}}\right) \qquad \hat{\sigma}_{ij} = \frac{1}{n-k}\left(\boldsymbol{Y}_{,i} - \boldsymbol{x}\hat{\boldsymbol{\theta}}_{,\boldsymbol{i}}\right)^T\left(\boldsymbol{Y}_{,j} - \boldsymbol{x}\hat{\boldsymbol{\theta}}_{,\boldsymbol{j}}\right)$$

$$V\left(\hat{\boldsymbol{\theta}}\right) = (\boldsymbol{x}^T\boldsymbol{x})^{-1} \otimes \Sigma; \qquad Cov\left(\hat{\boldsymbol{\theta}}_{,\boldsymbol{i}}, \hat{\boldsymbol{\theta}}_{,\boldsymbol{j}}\right) = \sigma_{ij}\left(\boldsymbol{x}^T\boldsymbol{x}\right)^{-1}$$

Multivariate General Linear Model V

The first problem is to estimate $\boldsymbol{\theta}$. We have

Theorem 4.14

We consider the above mentioned situation. If the observations $\boldsymbol{Y}_i$ are normally distributed the maximum likelihood estimate of $\boldsymbol{\theta}$ is given by

$$\hat{\boldsymbol{\theta}} = (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{Y}.$$

Remark 4.15

We see that

$$\hat{\boldsymbol{\theta}}_{j|} = (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{Y}_{j|},$$

i.e. the estimate for the j'th column in $\boldsymbol{\theta}$ is simply equal to the result we get by only considering the one dimensional general linear model for the j'th "property".

Example: The Skulls Data

```
> skulls<-read.csv2("Data/skulls2.csv")[,-1]
> skulls$epoch<-as.factor(paste(rep(1:5,each=30),skulls$epoch,sep="-"))

> summary(skulls)
       epoch          MB            BH              BL               NH
 1-c4000BC:30   Min.   :119   Min.   :120.0   Min.   : 81.00   Min.   :44.00
 2-c3300BC:30   1st Qu.:131   1st Qu.:129.0   1st Qu.: 93.00   1st Qu.:49.00
 3-c1850BC:30   Median :134   Median :133.0   Median : 96.00   Median :51.00
 4-c200BC :30   Mean   :134   Mean   :132.5   Mean   : 96.46   Mean   :50.93
 5-cAD150 :30   3rd Qu.:137   3rd Qu.:136.0   3rd Qu.:100.00   3rd Qu.:53.00
                Max.   :148   Max.   :145.0   Max.   :114.00   Max.   :60.00
```

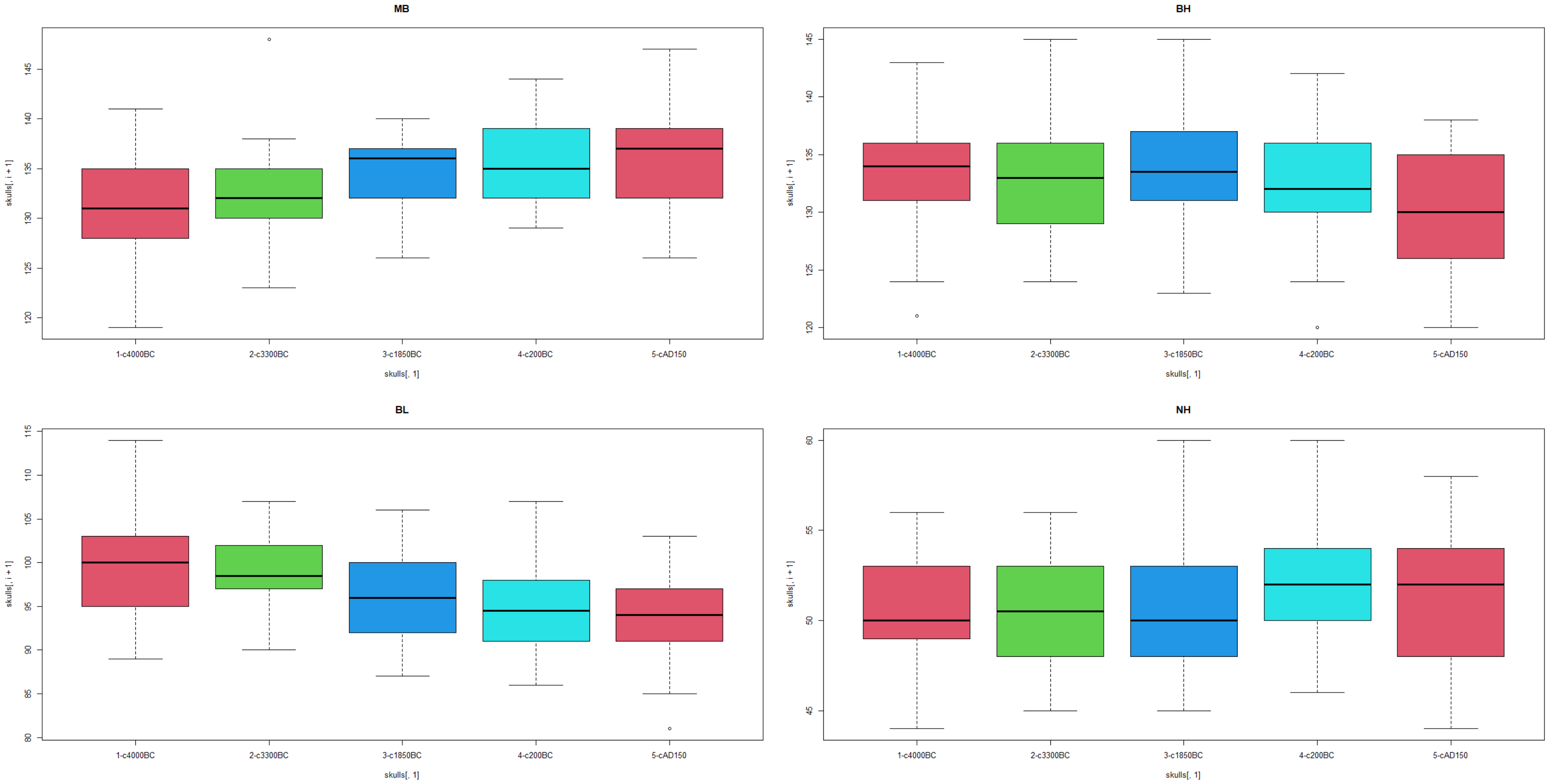
- Anthropometric measurements on male Egyptian skulls from 5 epochs (30 skulls from each epoch):

MB: Maximum Breath
BH: Basibregmatic Height
BL: Basialiveolar Length
NH: Nasal Length

Does the measurements evolve over time?
If so, how?

Analytical Method: Multivariate One-Way ANOVA

Example: The Skulls Data



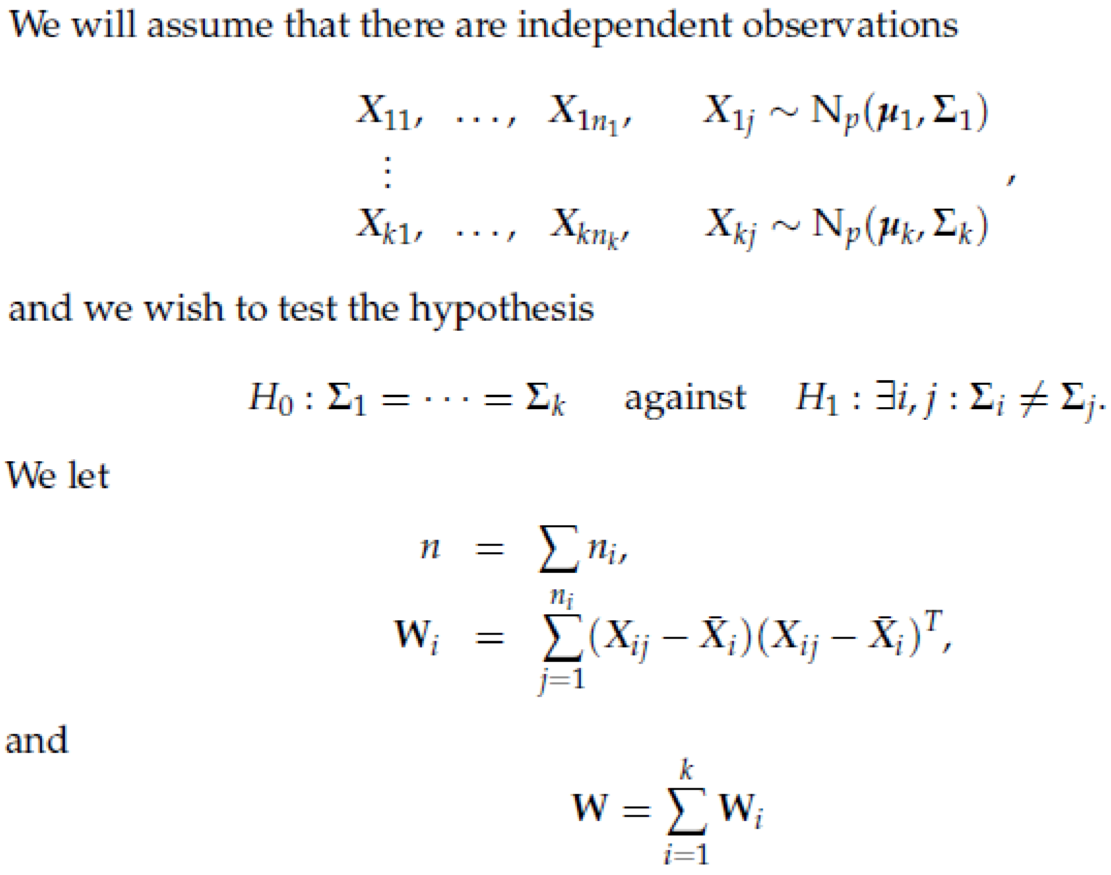
Example: The Skulls Data

- Let us first assume normality. If so, is the variance the same for each epoch?

- Univariate situation: Bartlett’s test.

- **Multivariate situation: Bartlett’s test.**

Example: The Skulls Data



We will assume that there are independent observations

$$\begin{array}{lll} X_{11}, \quad \ldots, & X_{1n_1}, & X_{1j} \sim \mathrm{N}_p(\boldsymbol{\mu}_1, \Sigma_1) \\ \vdots & & \\ X_{k1}, \quad \ldots, & X_{kn_k}, & X_{kj} \sim \mathrm{N}_p(\boldsymbol{\mu}_k, \Sigma_k) \end{array},$$

and we wish to test the hypothesis

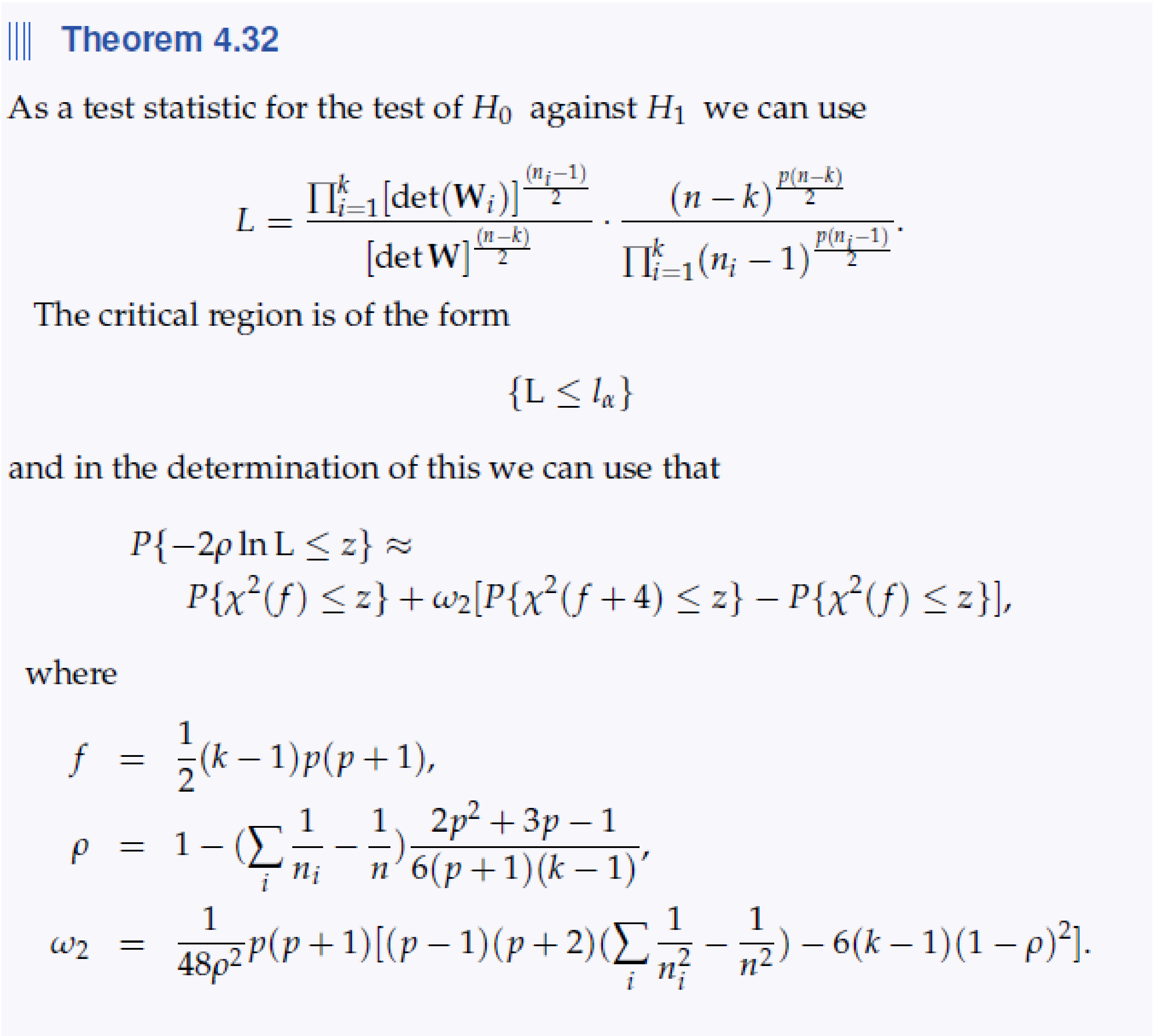
$$H_0 : \Sigma_1 = \cdots = \Sigma_k \quad \text{against} \quad H_1 : \exists i, j : \Sigma_i \neq \Sigma_j.$$

We let

$$\begin{aligned} n &= \sum n_i, \\ \mathbf{W}_i &= \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)^T, \end{aligned}$$

and

$$\mathbf{W} = \sum_{i=1}^{k} \mathbf{W}_i$$



Theorem 4.32

As a test statistic for the test of H_0 against H_1 we can use

$$L = \frac{\prod_{i=1}^{k} [\det(\mathbf{W}_i)]^{\frac{(n_i-1)}{2}}}{[\det \mathbf{W}]^{\frac{(n-k)}{2}}} \cdot \frac{(n-k)^{\frac{p(n-k)}{2}}}{\prod_{i=1}^{k} (n_i - 1)^{\frac{p(n_i-1)}{2}}}.$$

The critical region is of the form

$$\{\mathrm{L} \leq l_\alpha\}$$

and in the determination of this we can use that

$$\begin{aligned} &P\{-2\rho \ln \mathrm{L} \leq z\} \approx \\ &\quad P\{\chi^2(f) \leq z\} + \omega_2 [P\{\chi^2(f+4) \leq z\} - P\{\chi^2(f) \leq z\}], \end{aligned}$$

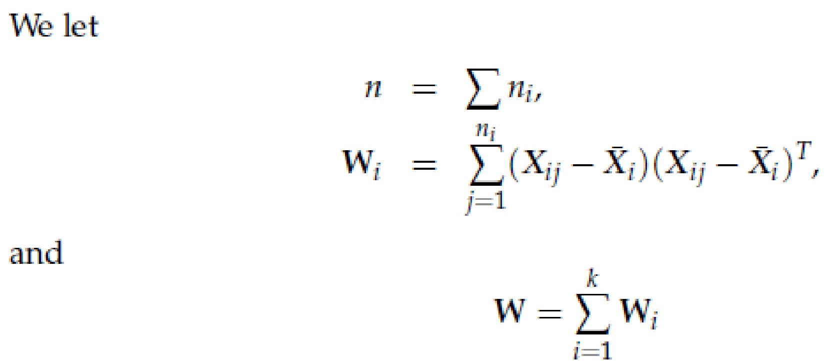
where

$$\begin{aligned} f &= \frac{1}{2}(k-1)p(p+1), \\ \rho &= 1 - \Big(\sum_i \frac{1}{n_i} - \frac{1}{n}\Big) \frac{2p^2 + 3p - 1}{6(p+1)(k-1)}, \\ \omega_2 &= \frac{1}{48\rho^2} p(p+1)\Big[(p-1)(p+2)\Big(\sum_i \frac{1}{n_i^2} - \frac{1}{n^2}\Big) - 6(k-1)(1-\rho)^2\Big]. \end{aligned}$$

Example: The Skulls Data

- calculating quantities for Bartlett's test

```
n<-150
n_i<-rep(30,5)
p<-4
k<-5

my.W<-list()
for(i in 1:5){
  temp<-residuals(manova(cbind(MB,BH,BL,NH)~1,
          data=skulls[skulls$epoch==levels(skulls$epoch)[i],]))
  my.W[[i]]<-t(temp)%*%temp
  }
W<-my.W[[1]]+my.W[[2]]+my.W[[3]]+my.W[[4]]+my.W[[5]]
```



We let

$$n = \sum n_i,$$
$$\mathbf{W}_i = \sum_{j=1}^{n_i} (\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)(\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)^T,$$

and

$$\mathbf{W} = \sum_{i=1}^{k} \mathbf{W}_i$$

Example: The Skulls Data

- incrementally building logL:

```
logL<-0
for(i in 1:5){logL<-logL+((n_i[i]-1)/2)*log(det(my.W[[i]]))}
logL<-logL-((n-k)/2)*log(det(W))
logL<-logL+(p*(n-k)/2)*log(n-k)
logL<-logL-sum((p*(n_i-1)/2)*log(n_i-1))
```

- Setting constants:

```
f<-(1/2)*(k-1)*p*(p+1)
rho<-1-(sum(1/n_i)-1/n)*(2*p^2+3*p-1)/(6*(p+1)*(k-1))
omega2<-(1/48)*p*(p+1)*((p-1)*(p+2)*(sum(1/n_i^2)-1/n^2)-6*(k-1)*(1-rho)^2)
```

- Test statistic:

```
(z<--2*rho*logL)
[1] 45.76321
```

- P-value:

```
# p-value:
(1-(pchisq(z,df=f)+omega2*(pchisq(z,df=f+4)-pchisq(z,df=f))))
[1] 0.2465424
```

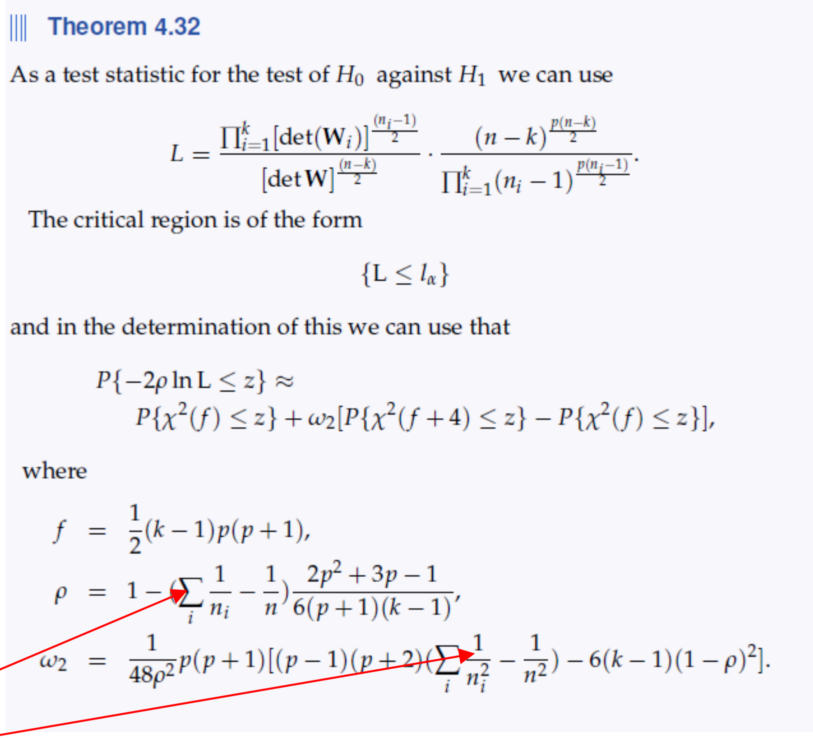
Example: The Skulls Data

- Test statistic `45.76321`, pvalue `0.2465424`.

- Test implemented in R in the `mbox` function:

```
> boxM(skulls[,2:5],skulls[,1])

        Box's M-test for Homogeneity of Covariance Matrices

data:  skulls[, 2:5]
Chi-Sq (approx.) = 45.667, df = 40, p-value = 0.2483
```

- Replace n_i with $n_i - 1$ and n with $n - k$ as they should be:
- And the test statistic will match **exactly** the one from `boxM`.



Theorem 4.32

As a test statistic for the test of H_0 against H_1 we can use

$$L = \frac{\prod_{i=1}^{k}[\det(\mathbf{W}_i)]^{\frac{(n_i-1)}{2}}}{[\det \mathbf{W}]^{\frac{(n-k)}{2}}} \cdot \frac{(n-k)^{\frac{p(n-k)}{2}}}{\prod_{i=1}^{k}(n_i-1)^{\frac{p(n_i-1)}{2}}}.$$

The critical region is of the form

$$\{\mathrm{L} \le l_\alpha\}$$

and in the determination of this we can use that

$$P\{-2\rho \ln \mathrm{L} \le z\} \approx P\{\chi^2(f) \le z\} + \omega_2[P\{\chi^2(f+4) \le z\} - P\{\chi^2(f) \le z\}],$$

where

$$\begin{aligned} f &= \frac{1}{2}(k-1)p(p+1), \\ \rho &= 1 - (\sum_i \frac{1}{n_i} - \frac{1}{n})\frac{2p^2+3p-1}{6(p+1)(k-1)}, \\ \omega_2 &= \frac{1}{48\rho^2}p(p+1)[(p-1)(p+2)(\sum_i \frac{1}{n_i^2} - \frac{1}{n^2}) - 6(k-1)(1-\rho)^2]. \end{aligned}$$

Example: The Skulls Data

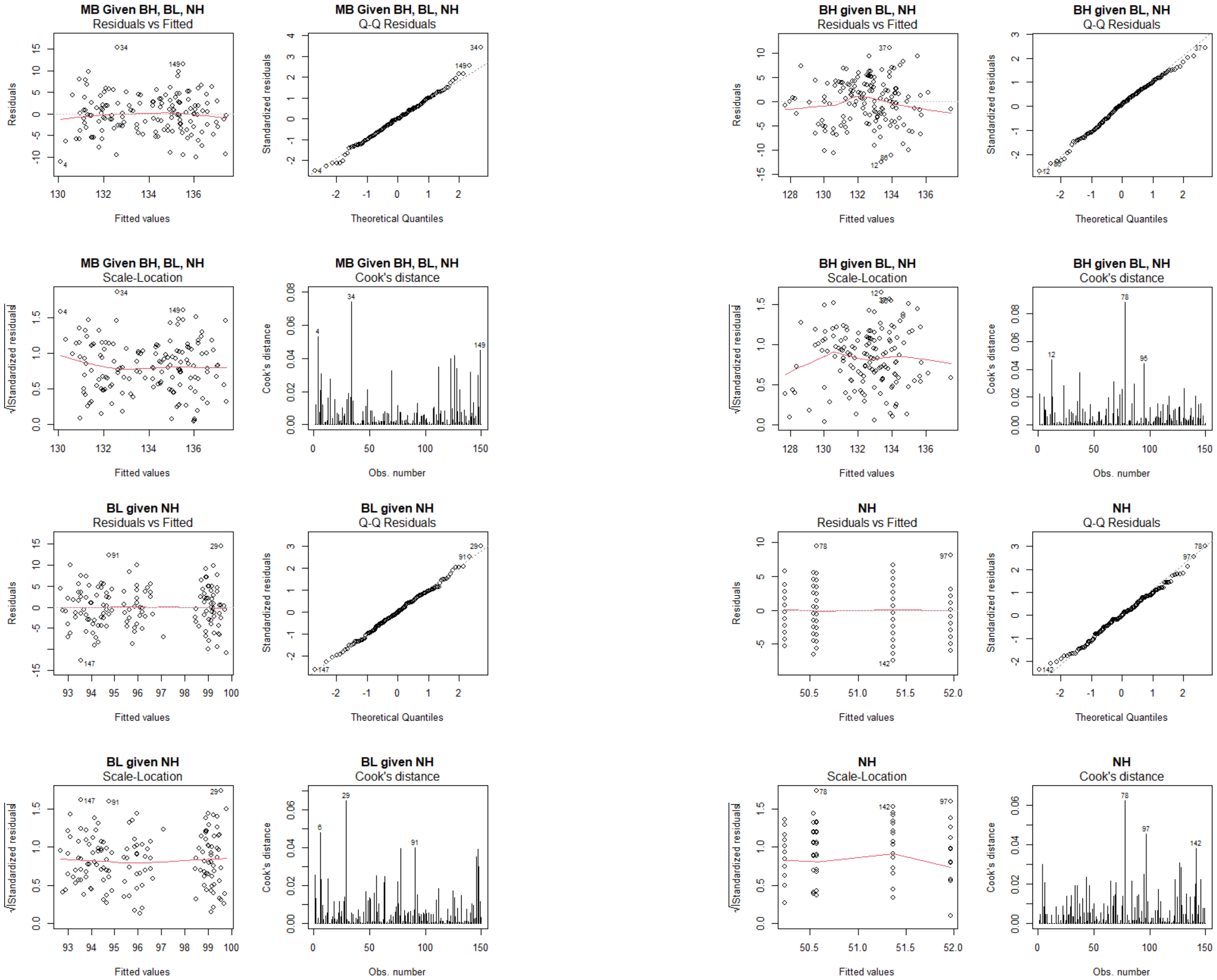
Model control (should be carried out for each epoch prior to Bartlett’s test, as in the R script):

Successive conditional distributions must be normal:

```
model.01<-lm(MB~epoch+BH+BL+NH,data=skulls)
model.02<-lm(BH~epoch+BL+NH,data=skulls)
model.03<-lm(BL~epoch+NH,data=skulls)
model.04<-lm(NH~epoch,data=skulls)
```

```
plot(model.01,which=1:4)
plot(model.02,which=1:4)
plot(model.03,which=1:4)
plot(model.04,which=1:4)
```

Example: The Skulls Data



Example: The Skulls Data

Simulataneous normality accepted.

$$Y = (MB \quad BH \quad BL \quad NH)$$

$$Y = X\theta + \varepsilon, \qquad \varepsilon \sim N_{150\times 4}(0, I_{150} \otimes \Sigma)$$

$$X_1 = \mathbf{1}\ ;$$
$$X_2 = 1_{\{c3300BC\}} - 1_{\{c4000BC\}};$$
$$X_3 = 1_{\{c1850BC\}} - 1_{\{c4000BC\}};$$
$$X_4 = 1_{\{c200BC\}} - 1_{\{c4000BC\}};$$
$$X_5 = 1_{\{cAD150\}} - 1_{\{c4000BC\}};$$

$$X = [X_1; X_2; X_3; X_4; X_5]$$

A note on row vs. column representation

- Why Y$\sim N_{150\times 4}(X\theta, I_{150}\otimes\Sigma)$ (each row is an observation) and not Y$\sim N_{4\times 150}(X\theta, \Sigma\otimes I_{150})$ (each column is an observation), like the book does through the vc function (for example page 290)?

- Semantics, but…

```
head(skulls)
      epoch  MB  BH  BL NH
1 1-c4000BC 131 138  89 49
2 1-c4000BC 125 131  92 48
3 1-c4000BC 131 132  99 50
4 1-c4000BC 119 132  96 44
5 1-c4000BC 136 143 100 54
6 1-c4000BC 138 137  89 56
```

- The form fits with standard spread-sheet represention of data.

Example: The Skulls Data

$$Y = X\theta + \varepsilon, \qquad \varepsilon \sim N_{150\times 4}(0, I_{150} \otimes \Sigma)$$

Corresponds to the hypothesis

$$H_0 : E(Y_i) = \theta_{epoch_i}, i = 1, \ldots, n.$$

Original purpose of study: Does the anthropometric measurements change over time?

Testing

$$H_1 : E(Y_i) = \theta \, , i = 1, \ldots, n.$$

Model:

$$Y = X\theta + \varepsilon, \qquad \varepsilon \sim N_{150\times 4}(0, I_{150} \otimes \Sigma)$$

With

$$X = \mathbf{1}$$

Example: The Skulls Data

Test statistic (lecture K):

$$R_1 = Q^{2/n} = \frac{det(SSD_0)}{det(SSD_0 + SSD_1)}$$

Where

$$SSD_0 = \sum_{epoch} \left(Y_{epoch} - \bar{Y}_{epoch}\right)^T \left(Y_{epoch} - \bar{Y}_{epoch}\right)$$

$$SSD_1 = 30 \sum_{epoch} \left(\bar{Y}_{epoch} - \bar{Y}\right)^T \left(\bar{Y}_{epoch} - \bar{Y}\right)$$

since there are 30 measurements for each epoch.

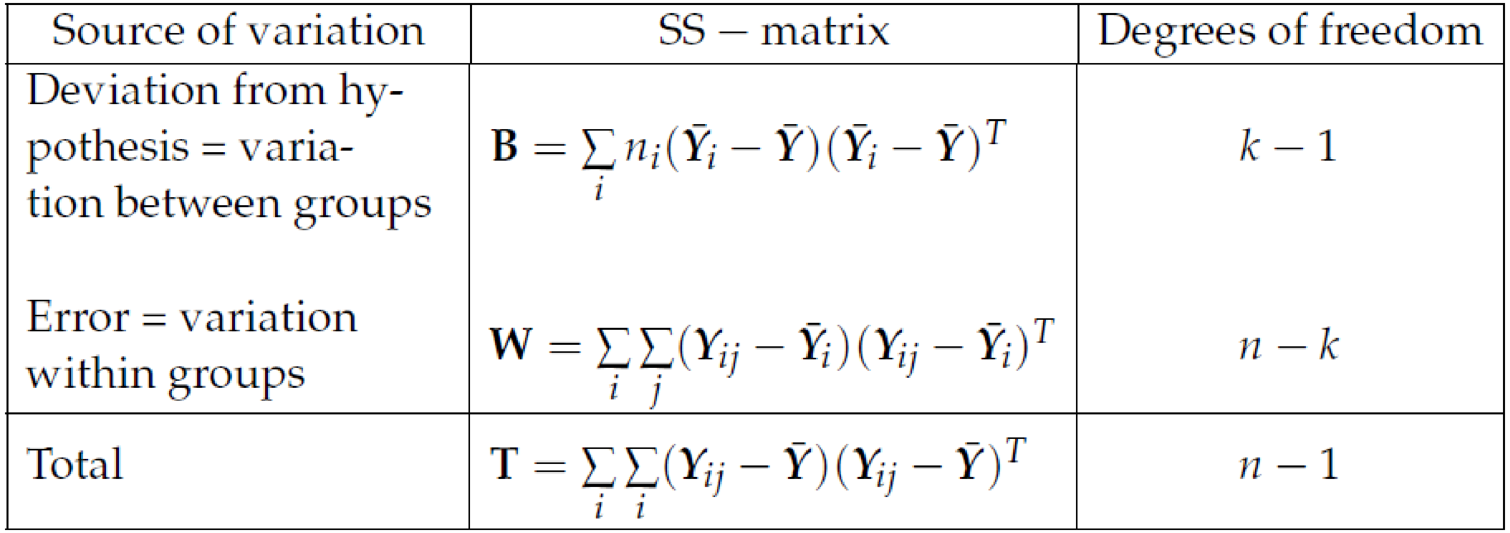
- SSD_0 is the variation within groups: W.
- SSD_1 is the variation between groups: B.
- If SSD_1 is small, R_1 is close to 1.

Example: The Skulls Data

Decomposition of Total Variation:

Total Variation= Variation Within Groups + Variation between groups:

$$SSD_{01} = SSD_0 + SSD_1$$
$$T = W + B$$



Source of variation	SS – matrix	Degrees of freedom
Deviation from hypothesis = variation between groups	$\mathbf{B} = \sum_i n_i(\bar{\boldsymbol{Y}}_i - \bar{\boldsymbol{Y}})(\bar{\boldsymbol{Y}}_i - \bar{\boldsymbol{Y}})^T$	$k-1$
Error = variation within groups	$\mathbf{W} = \sum_i \sum_j (\boldsymbol{Y}_{ij} - \bar{\boldsymbol{Y}}_i)(\boldsymbol{Y}_{ij} - \bar{\boldsymbol{Y}}_i)^T$	$n-k$
Total	$\mathbf{T} = \sum_i \sum_i (\boldsymbol{Y}_{ij} - \bar{\boldsymbol{Y}})(\boldsymbol{Y}_{ij} - \bar{\boldsymbol{Y}})^T$	$n-1$

Example: The Skulls Data

- Testing:

```
> analysis<-manova(cbind(MB,BH,BL,NH)~epoch,data=skulls)

> summary(analysis,test="Wilks")
           Df   Wilks approx F num Df den Df    Pr(>F)
epoch       4 0.66359   3.9009     16 434.45 7.01e-07 ***
Residuals 145
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- The data support changes over time of anthropometric measurements (p<0.001).

Example: The Skulls Data

- Other test statistics:

```
> summary(analysis)
           Df  Pillai approx F num Df den Df    Pr(>F)    
epoch       4 0.35331    3.512     16    580 4.675e-06 ***
Residuals 145
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(analysis,test="Hotelling-Lawley")
           Df Hotelling-Lawley approx F num Df den Df    Pr(>F)    
epoch       4          0.48182    4.231     16    562 8.278e-08 ***
Residuals 145
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(analysis,test="Roy")
           Df    Roy approx F num Df den Df    Pr(>F)    
epoch       4 0.4251    15.41      4    145 1.588e-10 ***
Residuals 145
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Pillai´s Trace test, the Hotelling-Lawley test, Roy’s largest root test.

- Pillai´s Trace test is the default in the manova summary. You need to specify the Wilks test to get it.

- For definitions, see remark 4.23.

- Each of these other tests have their advantages, in terms of a higher power towards specific alternatives, or less sensitivitiy from departures from normality or variance homogeneity.

- **We shall not explore these concepts**,, but stick to **the Wilks test**, which has the advantage that it is **equivalent to the Likelihood Ratio test**, as we have seen.

Example: The Skulls Data

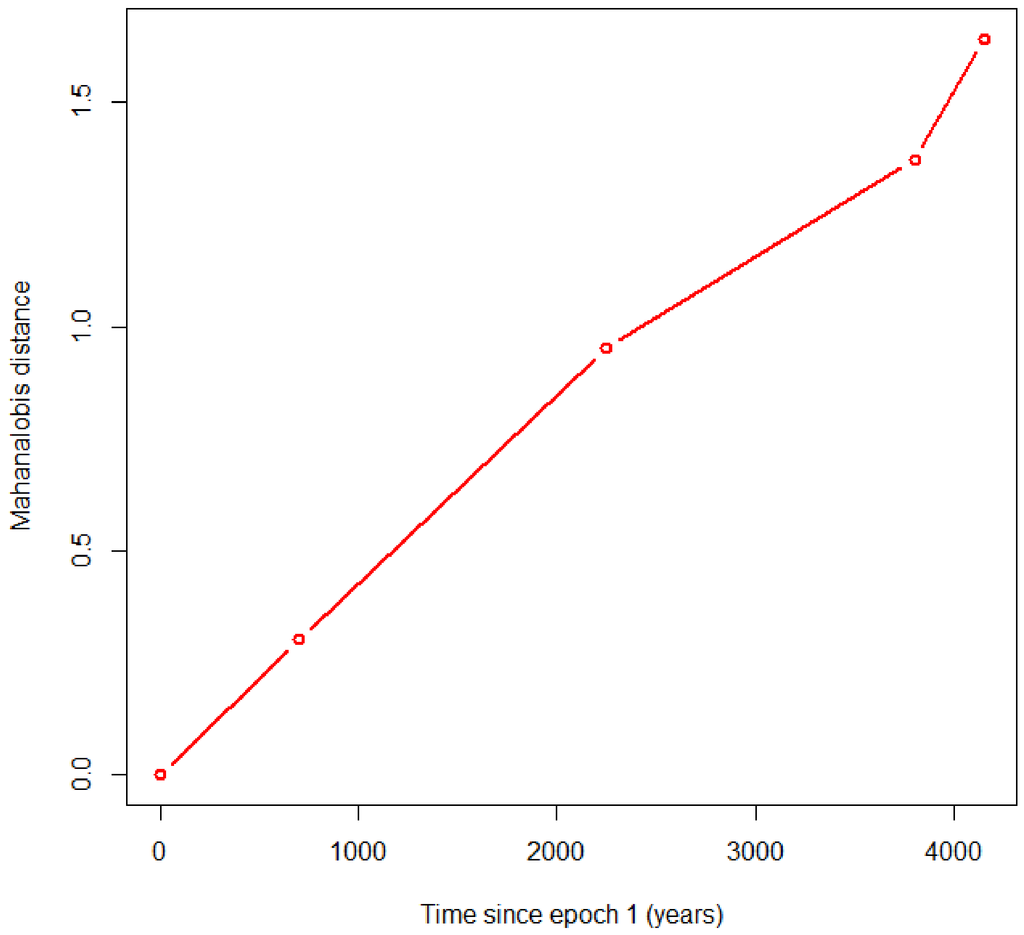
- Post hoc analysis: Re-parametrize to get levels within each epoch:

```
> (theta<-coef(manova(cbind(MB,BH,BL,NH)~epoch-1,data=skulls)))
                     MB       BH       BL       NH
epoch1-c4000BC 131.3667 133.6000 99.16667 50.53333
epoch2-c3300BC 132.3667 132.7000 99.06667 50.23333
epoch3-c1850BC 134.4667 133.8000 96.03333 50.56667
epoch4-c200BC  135.5000 132.3000 94.53333 51.96667
epoch5-cAD150  136.1667 130.3333 93.50000 51.36667

> R<-residuals(manova(cbind(MB,BH,BL,NH)~as.factor(epoch),data=skulls))
> (Sigma<-(1/(150-5))*t(R)%*%R)
            MB          BH          BL        NH
MB 21.11080460  0.03678161  0.07908046  2.008966
BH  0.03678161 23.48459770  5.20000000  2.845057
BL  0.07908046  5.20000000 24.17908046  1.133333
NH  2.00896552  2.84505747  1.13333333 10.152644
```

Example: The Skulls Data

- Mahanalobis distance to epoch1:

$$\left\|\hat{\theta}_i - \hat{\theta}_1\right\|_{\hat{\Sigma}^{-1}},\ i = 1, \ldots, 5$$

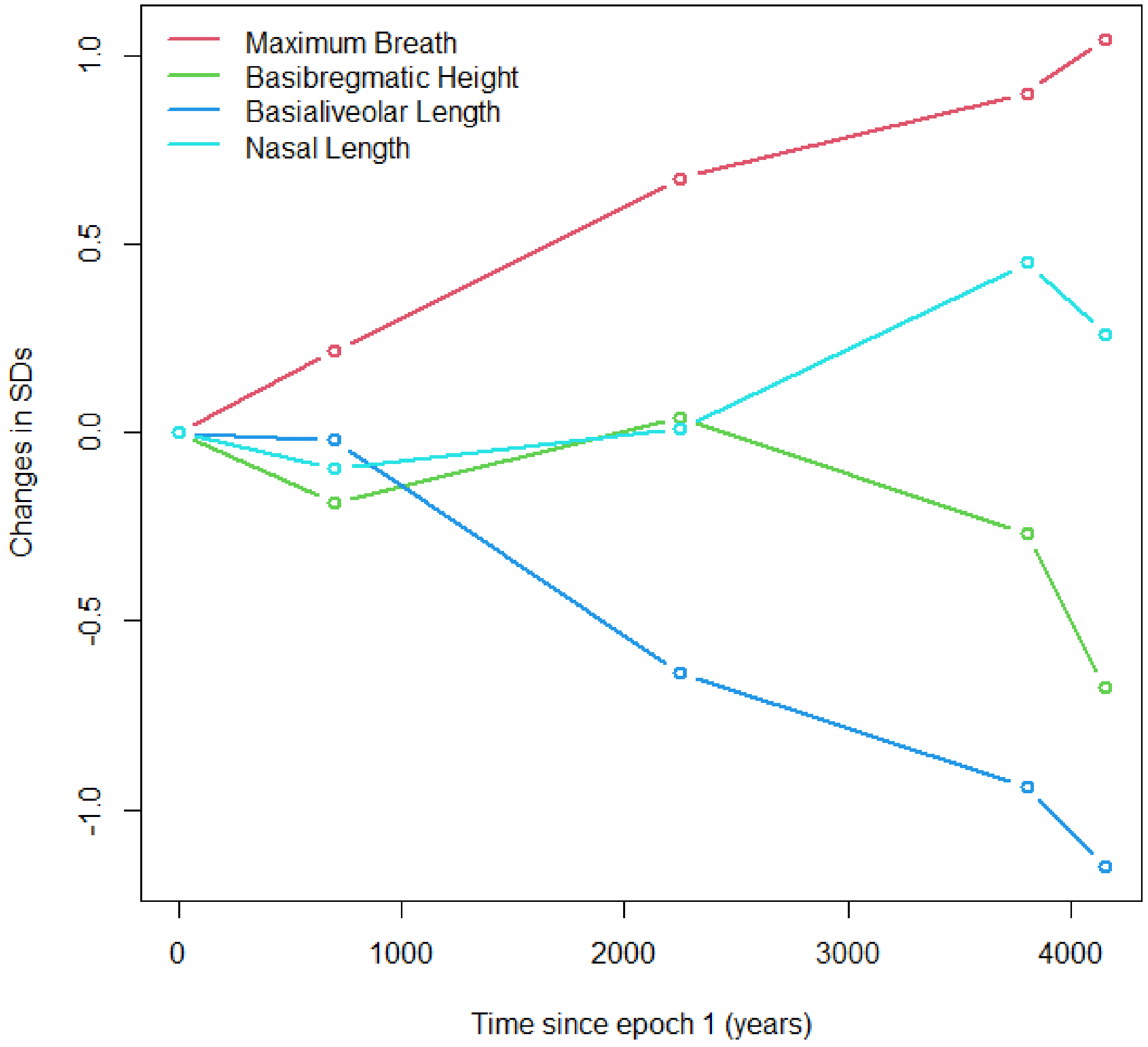


- A more or less linear progression in changes. Makes sense through constant genetic drift.

Example: The Skulls Data

- Changes of individual measurements:

- Changes towards wider and less protruding skulls (basialiveolar length).



Two-Way Multivariate Analysis of Variance

- **Hypotheses in two-way ANOVA (low birthweight example, Lecture J):**

H_2: Additive hypothesis ($\delta_{rs} = 0$) :

$$M_2\colon \mu_{rsi} = \alpha + \beta_r + \gamma_s$$

H_{31}: No effect of smoking ($\gamma_s = 0$):

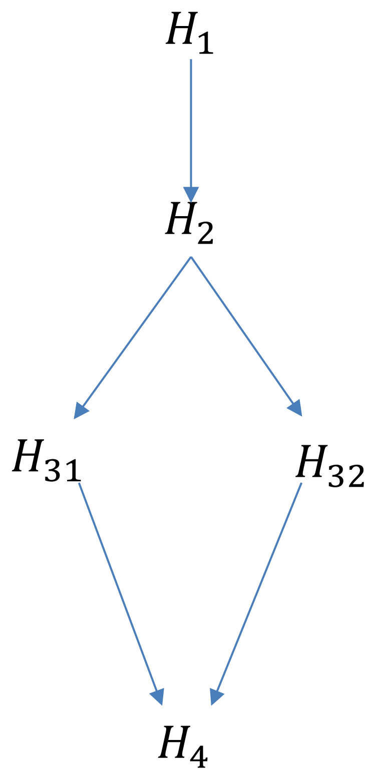
$$M_{31}\colon \mu_{rsi} = \alpha + \beta_r$$

H_{32}: No effect of race ($\beta_r = 0$):

$$M_{32}\colon \mu_{rsi} = \alpha + \gamma_s$$

H_4: No effect of neither smoking nor race ($\beta_r = 0$ or $\gamma_s = 0$):

$$M_4\colon \mu_{rsi} = \alpha$$



H_1

H_2

H_{31}

H_{32}

H_4

Two-Way Multivariate Analysis of Variance

- Assume that the mean of Y is affected by two factors.

- If each combination is only observed once, the interaction model

$$M_1\colon Y_{ij} = \mu_{ij} + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim N_p(0, \Sigma), i = 1, \dots, k, j = 1, \dots, m$$

is **not testable**; often the case in multivariate analysis.

- We shall start at the additive model:

$$M_2\colon Y_{ij} = \alpha + \beta_i + \gamma_j + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim N_p(0, \Sigma)$$

We do not make the assumptions in the book that $\sum\beta_i = \sum\gamma_j = 0$; such assumptions belong to a particular parametrization, which we do not need to concern ourselves with here.

- We have two hypotheses to test:

H_{31}: No effect of the second factor ($\gamma_j = 0$):

$$M_{31}\colon \mu_{ij} = \alpha + \beta_i$$

H_{32}: No effect of the first factor ($\beta_r = 0$):

$$M_{32}\colon \mu_{ij} = \alpha + \gamma_j$$

Two-Way Multivariate Analysis of Variance

$$M_2\colon Y_{ij} = \alpha + \beta_i + \gamma_j + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim N_p(0, \Sigma)$$

- H_1, H_2 subspaces of a vector space;

$$H_1 \otimes H_2 = \{y_1 + y_2 | y_1 \in \mathrm{H}_1, y_2 \in H_2\}$$

- The projection on a direct sum of two geometric orthogonal subspaces $H_1 \otimes H_2$ can be found as the sum of the projections on each of them, and then subtracting the projection on the intersection:

$$P_{H_1 \otimes H_2} = P_{H_1} + P_{H_2} - P_{H_1 \cap H_2}$$

$$\dim(H_1 \otimes H_2) = \dim(H_1) + \dim(H_2) - \dim(H_1 \cap H_2)$$

- Thus, the SSD corresponding to M_2 is seen to be

$$SSD_{02} = \sum_{i=1}^{k} \sum_{j=1}^{m} \left(Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{\cdot\cdot}\right)^T \left(Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{\cdot\cdot}\right)$$

Two-Way Multivariate Analysis of Variance

$$M_2 : Y_{ij} = \alpha + \beta_i + \gamma_j + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim N_p(0, \Sigma)$$

- Variations between 1st factor levels, and variations between 2nd factor levels:

$$SSD_{31} = \sum_{i=1}^{k} \sum_{j=1}^{m} \left(\left(\bar{Y}_{i\cdot} + \bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot} \right) - \bar{Y}_{i\cdot} \right)^T \left(\left(\bar{Y}_{i\cdot} + \bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot} \right) - \bar{Y}_{i\cdot} \right) = k \sum_{j=1}^{m} \left(\bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot} \right)^T \left(\bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot} \right)$$

$$SSD_{32} = \sum_{i=1}^{k} \sum_{j=1}^{m} \left(\left(\bar{Y}_{i\cdot} + \bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot} \right) - \bar{Y}_{\cdot j} \right)^T \left(\left(\bar{Y}_{i\cdot} + \bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot} \right) - \bar{Y}_{\cdot j} \right) = m \sum_{i=1}^{k} \left(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot} \right)^T \left(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot} \right)$$

Two-Way Multivariate Analysis of Variance

$$M_2 : Y_{ij} = \alpha + \beta_i + \gamma_j + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim N_p(0, \Sigma)$$

- Splitting the total variation:

$$Y_{ij} - \bar{Y}_{..} = \left(Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..}\right) + \left(\bar{Y}_{.j} - \bar{Y}_{..}\right) + (\bar{Y}_{i.} - \bar{Y}_{..})$$

$$\left\|Y_{ij} - \bar{Y}_{..}\right\|^2 = \left\|Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..}\right\|^2 + \left\|\bar{Y}_{.j} - \bar{Y}_{..}\right\|^2 + \left\|\bar{Y}_{i.} - \bar{Y}_{..}\right\|^2$$

$$SSD_0 \qquad = \qquad SSD_{02} \qquad\qquad + \quad SSD_{31} \quad + \quad SSD_{32}$$

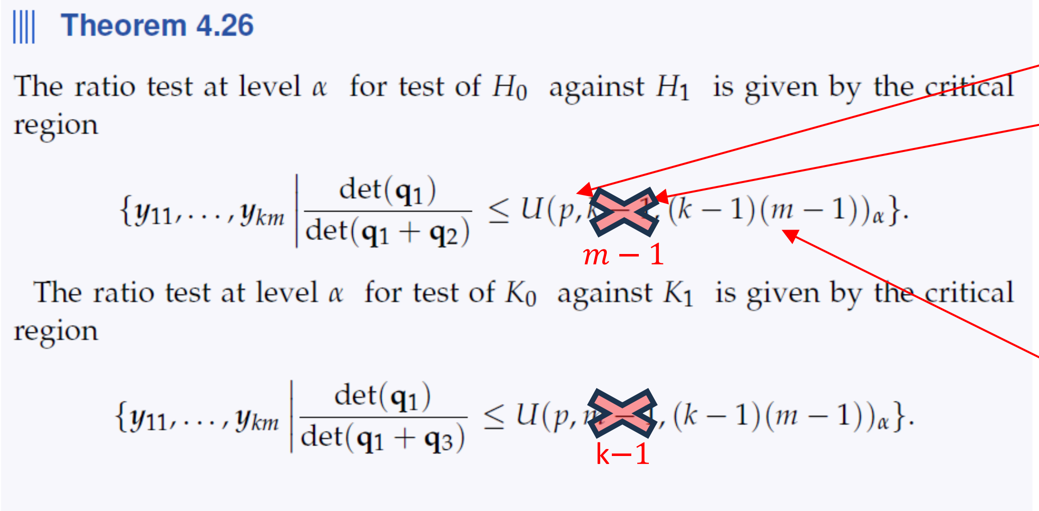
Two-Way Multivariate Analysis of Variance

$$M_2: Y_{ij} = \alpha + \beta_i + \gamma_j + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim N_p(0, \Sigma)$$

- Note that the subspace H_2 corresponding to the additive hypothesis (here) has dimension $k + m - 1$. $\dim(H_{31}) = k, \dim(H_{32}) = m$.
- Also note that $n - \dim(H_2) = km - (k + m - 1) = (k - 1)(m - 1)$.

- Test statistics:

$$R_{31} = \frac{\det(SSD_{02})}{det(SSD_{02} + SSD_{31})}, R_{32} = \frac{\det(SSD_{02})}{det(SSD_{02} + SSD_{32})}$$



Theorem 4.26

The ratio test at level α for test of H_0 against H_1 is given by the critical region

$$\left\{ y_{11}, \ldots, y_{km} \,\middle|\, \frac{\det(\mathbf{q}_1)}{\det(\mathbf{q}_1 + \mathbf{q}_2)} \le U(p, \underset{m-1}{\cancel{k-1}}, (k-1)(m-1))_\alpha \right\}.$$

The ratio test at level α for test of K_0 against K_1 is given by the critical region

$$\left\{ y_{11}, \ldots, y_{km} \,\middle|\, \frac{\det(\mathbf{q}_1)}{\det(\mathbf{q}_1 + \mathbf{q}_3)} \le U(p, \underset{k-1}{\cancel{m-1}}, (k-1)(m-1))_\alpha \right\}.$$

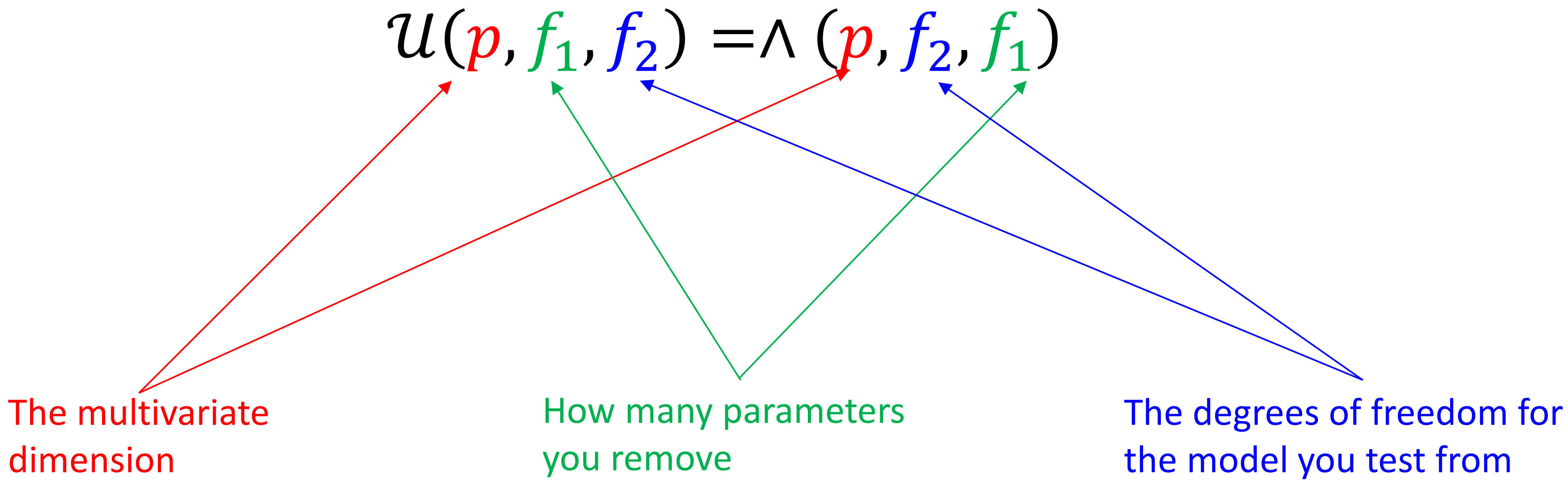
The multivariate dimension

How many parameters you remove

The degrees of freedom for the model you test from

Two-Way Multivariate Analysis of Variance

- Important note on the books notation through Anderson's $\mathcal{U}$:



$$\mathcal{U}(p, f_1, f_2) = \Lambda\,(p, f_2, f_1)$$

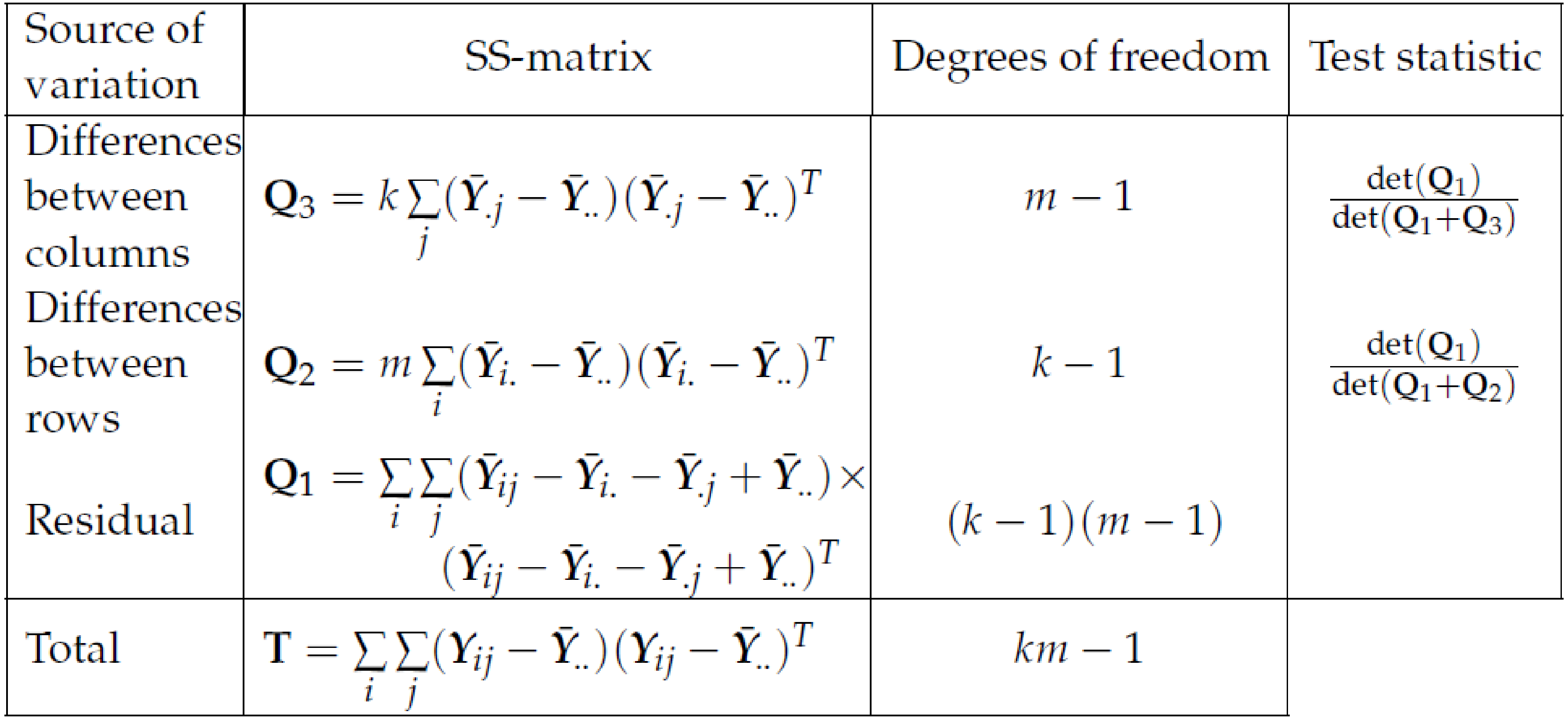
The multivariate dimension

How many parameters you remove

The degrees of freedom for the model you test from

- In practical applications the distinction is of no importance, as the software will detect the parameter values intrinsic.

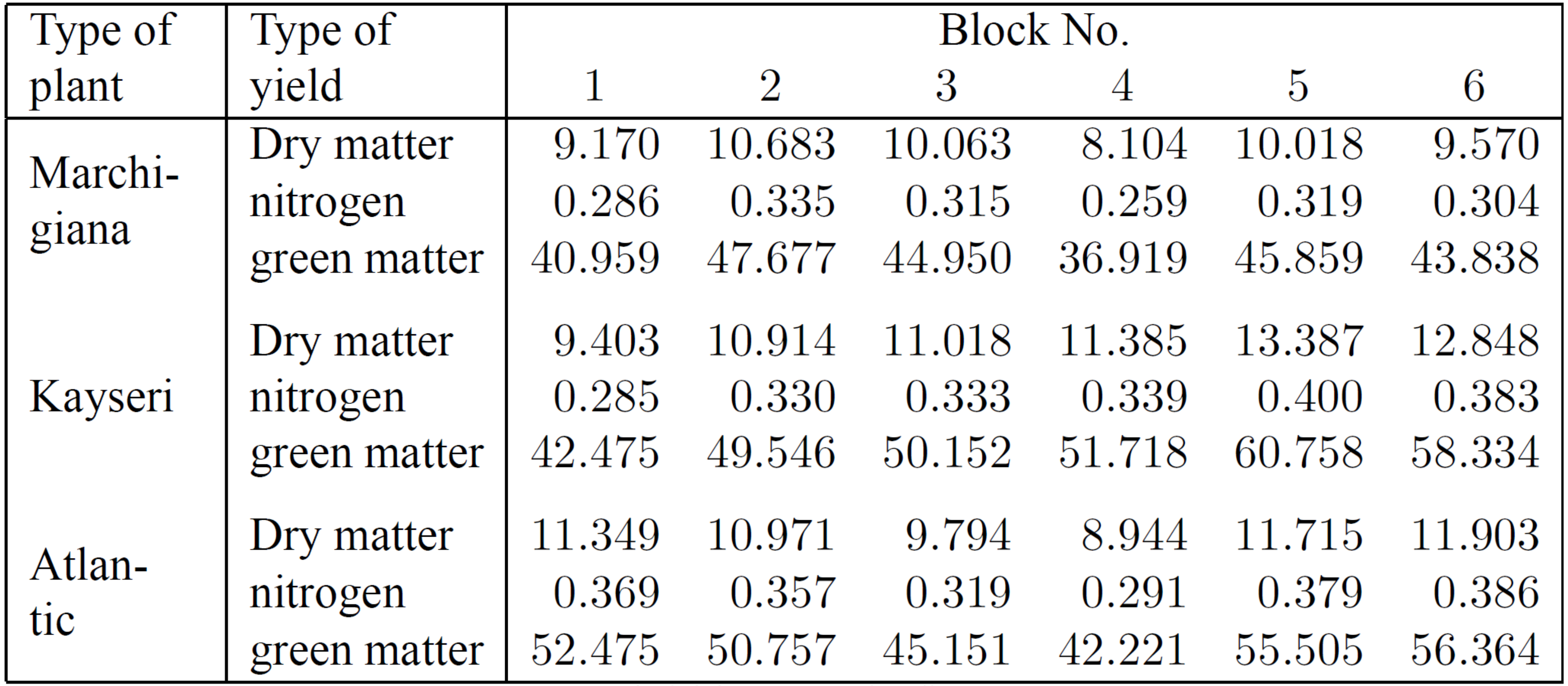
Two-Way Multivariate Analysis of Variance



Source of variation	SS-matrix	Degrees of freedom	Test statistic
Differences between columns	$\mathbf{Q}_3 = k\sum_j(\bar{\mathbf{Y}}_{.j} - \bar{\mathbf{Y}}_{..})(\bar{\mathbf{Y}}_{.j} - \bar{\mathbf{Y}}_{..})^T$	$m - 1$	$\frac{\det(\mathbf{Q}_1)}{\det(\mathbf{Q}_1+\mathbf{Q}_3)}$
Differences between rows	$\mathbf{Q}_2 = m\sum_i(\bar{\mathbf{Y}}_{i.} - \bar{\mathbf{Y}}_{..})(\bar{\mathbf{Y}}_{i.} - \bar{\mathbf{Y}}_{..})^T$	$k - 1$	$\frac{\det(\mathbf{Q}_1)}{\det(\mathbf{Q}_1+\mathbf{Q}_2)}$
Residual	$\mathbf{Q}_1 = \sum_i\sum_j(\bar{\mathbf{Y}}_{ij} - \bar{\mathbf{Y}}_{i.} - \bar{\mathbf{Y}}_{.j} + \bar{\mathbf{Y}}_{..})\times$ $(\bar{\mathbf{Y}}_{ij} - \bar{\mathbf{Y}}_{i.} - \bar{\mathbf{Y}}_{.j} + \bar{\mathbf{Y}}_{..})^T$	$(k-1)(m-1)$	
Total	$\mathbf{T} = \sum_i\sum_j(\mathbf{Y}_{ij} - \bar{\mathbf{Y}}_{..})(\mathbf{Y}_{ij} - \bar{\mathbf{Y}}_{..})^T$	$km - 1$	

- Unbiased estimator for Σ: $\hat{\Sigma} = \frac{1}{(k-1)(m-1)} SSD_{02}, \hat{\Sigma} = \frac{1}{(k-1)(m-1)} Q_1$

Example: Plant Yield



Type of plant	Type of yield	Block No.					
		1	2	3	4	5	6
Marchigiana	Dry matter	9.170	10.683	10.063	8.104	10.018	9.570
	nitrogen	0.286	0.335	0.315	0.259	0.319	0.304
	green matter	40.959	47.677	44.950	36.919	45.859	43.838
Kayseri	Dry matter	9.403	10.914	11.018	11.385	13.387	12.848
	nitrogen	0.285	0.330	0.333	0.339	0.400	0.383
	green matter	42.475	49.546	50.152	51.718	60.758	58.334
Atlantic	Dry matter	11.349	10.971	9.794	8.944	11.715	11.903
	nitrogen	0.369	0.357	0.319	0.291	0.379	0.386
	green matter	52.475	50.757	45.151	42.221	55.505	56.364

Yield in 1000 kg/ha

- Case 4.27 in the book.
- Plots are soil of different quality.

Example: Plant Yield

- One observation per combination of plant type and block.
- 3 responses (type of yield): *dry matter, nitrogen, green matter*.

- Model:

$$Y_{ij} = \begin{pmatrix} dry\ matter \\ nitrogen \\ green\ matter \end{pmatrix}_{\mathrm{ij}} = \alpha + \beta_{type(i)} + \gamma_{block(j)} + \varepsilon_{ij}, \varepsilon_{ij} \sim N_3(0, \Sigma), i = 1, \dots, 3, j = 1, \dots, 6.$$

Example: Plant Yield

```
> yield<-read.table("Data/plant yield.txt",header=T)
> yield$Type<-as.factor(yield$Type)
> yield$Yield<-as.factor(yield$Yield)

> summary(yield)
         Type            Yield         B1               B2               B3               B4               B5               B6
Atlantic   :3   Dry matter  :3   Min.   : 0.285   Min.   : 0.330   Min.   : 0.315   Min.   : 0.259   Min.   : 0.319   Min.   : 0.304
Kayseri    :3   Green matter:3   1st Qu.: 0.369   1st Qu.: 0.357   1st Qu.: 0.333   1st Qu.: 0.339   1st Qu.: 0.400   1st Qu.: 0.386
Marchigiana:3   Nitrogen    :3   Median : 9.403   Median :10.914   Median :10.063   Median : 8.944   Median :11.715   Median :11.903
                                 Mean   :18.530   Mean   :20.174   Mean   :19.122   Mean   :17.798   Mean   :22.038   Mean   :21.548
                                 3rd Qu.:40.959   3rd Qu.:47.677   3rd Qu.:44.950   3rd Qu.:36.919   3rd Qu.:45.859   3rd Qu.:43.838
                                 Max.   :52.475   Max.   :50.757   Max.   :50.152   Max.   :51.718   Max.   :60.758   Max.   :58.334
```

Example: Plant Yield

- One observation per combination of plant type and block.
- 3 responses (type of yield): *dry matter, nitrogen, green matter*.

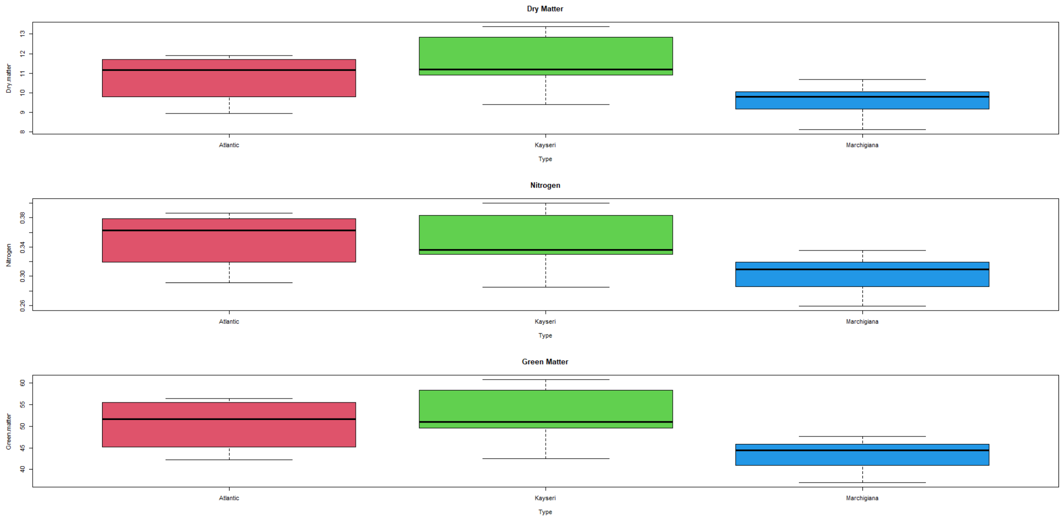
- Model:

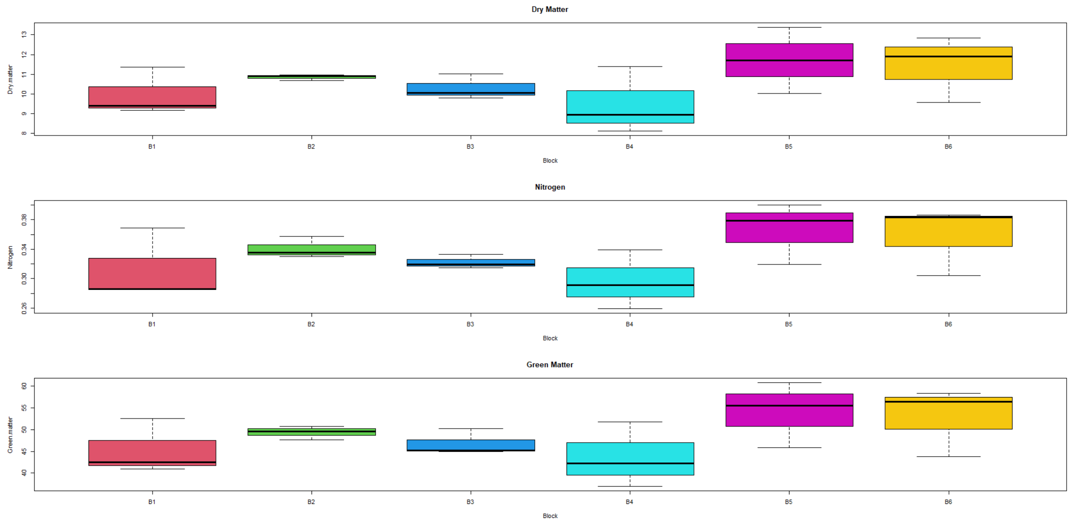
$$Y_{ij} = \begin{pmatrix} dry\ matter \\ nitrogen \\ green\ matter \end{pmatrix}_{\mathrm{ij}} = \alpha + \beta_{type(i)} + \gamma_{block(j)} + \varepsilon_{ij}, \varepsilon_{ij} \sim N_3(0, \Sigma), i = 1, \dots, 3, j = 1, \dots, 6.$$

Example: Plant Yield

- Recoding to fit MANOVA: Each row needs to be an observation.

```
> yield2$Block<-rep(paste("B",1:6,sep=""),3)
> yield2$Dry.matter<-c(t(yield[c(1,4,7),-(1:2)]))
> yield2$Nitrogen<-c(t(yield[c(2,5,8),-(1:2)]))
> yield2$Green.matter<-c(t(yield[c(3,6,9),-(1:2)]))
> yield2
         Type Block Dry.matter Nitrogen Green.matter
1  Marchigiana    B1      9.170    0.286       40.959
2  Marchigiana    B2     10.683    0.335       47.677
3  Marchigiana    B3     10.063    0.315       44.950
4  Marchigiana    B4      8.104    0.259       36.919
5  Marchigiana    B5     10.018    0.319       45.859
6  Marchigiana    B6      9.570    0.304       43.838
7      Kayseri    B1      9.403    0.285       42.475
8      Kayseri    B2     10.914    0.330       49.546
9      Kayseri    B3     11.018    0.333       50.152
10     Kayseri    B4     11.385    0.339       51.718
11     Kayseri    B5     13.387    0.400       60.758
12     Kayseri    B6     12.848    0.383       58.334
13    Atlantic    B1     11.349    0.369       52.475
14    Atlantic    B2     10.971    0.357       50.757
15    Atlantic    B3      9.794    0.319       45.151
16    Atlantic    B4      8.944    0.291       42.221
17    Atlantic    B5     11.715    0.379       55.505
18    Atlantic    B6     11.903    0.386       56.364
```

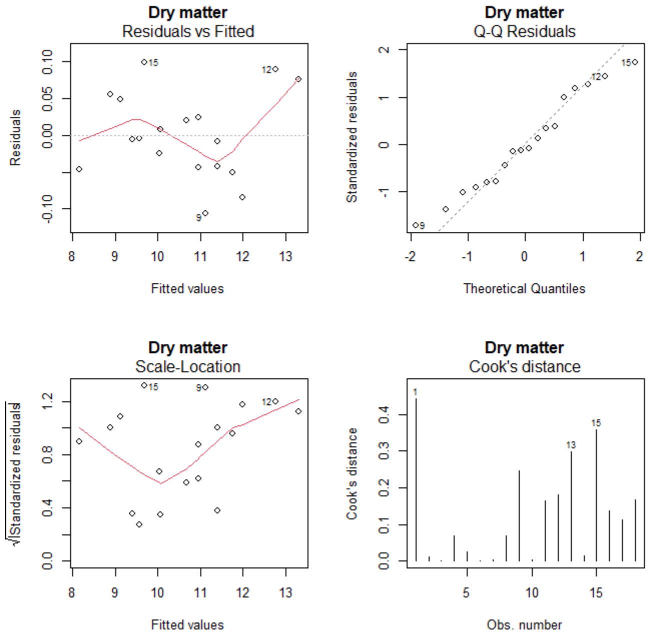


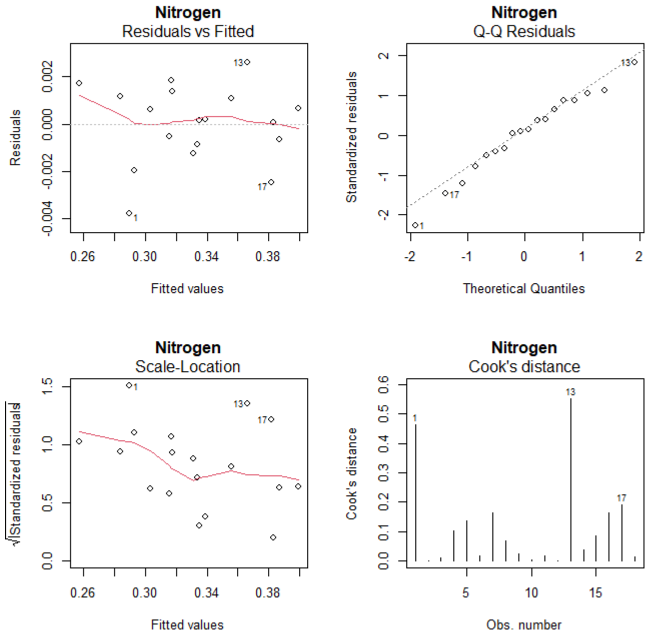


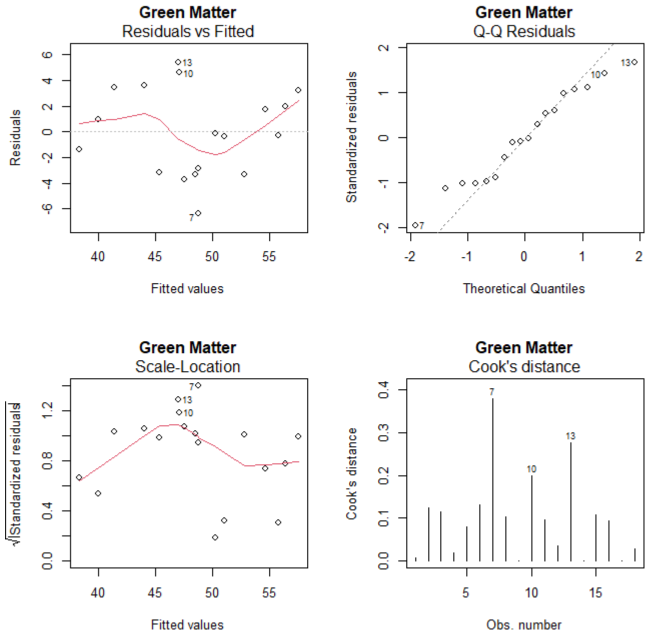
Example: Plant Yield

- Model control:

```
analysis.01<-lm(Dry.matter~Type+Block+Nitrogen+Green.matter,data=yield2)
analysis.02<-lm(Nitrogen~Type+Block+Green.matter,data=yield2)
analysis.03<-lm(Green.matter~Type+Block,data=yield2)
```







Example: Plant Yield

- Looking at a somewhat influential observation:

```
> summary(yield2)
     Type              Block             Dry.matter        Nitrogen       Green.matter
 Length:18          Length:18          Min.   : 8.104   Min.   :0.2590   Min.   :36.92
 Class  :character  Class  :character  1st Qu.: 9.626   1st Qu.:0.3068   1st Qu.:44.12
 Mode  :character   Mode  :character   Median :10.799   Median :0.3315   Median :48.61
                                       Mean   :10.624   Mean   :0.3327   Mean   :48.65
                                       3rd Qu.:11.376   3rd Qu.:0.3660   3rd Qu.:52.29
                                       Max.   :13.387   Max.   :0.4000   Max.   :60.76
> yield2[13,]
       Type Block Dry.matter Nitrogen Green.matter
13 Atlantic    B1     11.349    0.369       52.475
```

- Observation 13 does not seem to be extreme in any response.
- Does not indicate a problematic observation.

Example: Plant Yield

- Two-way MANOVA:

```
> analysis<-manova(cbind(Dry.matter,Nitrogen,Green.matter)~Type+Block,data=yield2)

> summary(analysis,test="Wilks")
          Df    Wilks approx F num Df den Df    Pr(>F)
Type       2 0.003326   43.573      6 16.000 4.956e-09 ***
Block      5 0.063013    2.581     15 22.486   0.02055 *
Residuals 10
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Both Type of plant (p<0.001) and Block (p=0.02) has a significant impact on the yield.

Example: Plant Yield

- MANOVA Table:

Source of variation	SSD	df	Test
Block	Dry.matter Nitrogen Green.matter Dry.matter 11.2182476 0.34801194 53.974649 Nitrogen 0.3480119 0.01080228 1.671297 Green.matter 53.9746487 1.67129667 261.702513	$6-1=5$	0.063013 $\sim \Lambda(3,(3-1)(5-1),6-1)$ $=\Lambda(3,10,5)$ $p=0.02$
Type	Dry.matter Nitrogen Green.matter Dry.matter 10.9456041 0.262613278 52.369429 Nitrogen 0.2626133 0.008030778 1.385427 Green.matter 52.3694292 1.385427500 260.173972	$3-1=2$	0.003326 $\sim \Lambda(3,(3-1)(5-1),3-1)$ $=\Lambda(3,10,2)$ $p<0.0001$
Residuals	Dry.matter Nitrogen Green.matter Dry.matter 9.9701146 0.286666722 43.454086 Nitrogen 0.2866667 0.008310556 1.255111 Green.matter 43.4540855 1.255111167 190.600538	$(3-1)(6-1)=10$	
Total	Dry.matter Nitrogen Green.matter Dry.matter 32.1339663 0.89729194 149.798163 Nitrogen 0.8972919 0.02714361 4.311835 Green.matter 149.7981633 4.31183533 712.477024	$6*3-1=17$	

- Residuals mean less for Green matter; more variation is systematic.
- For Dry matter and Nitrogen, Block, Plant Type and Residual variation contributes equally.

Example: Plant Yield

- Univariate analyses:

```
> analysis.dm<-lm(Dry.matter~Type+Block,data=yield2)
> analysis.ni<-lm(Nitrogen~Type+Block,data=yield2)
> analysis.gm<-lm(Green.matter~Type+Block,data=yield2)
> drop1(analysis.dm,test="F")

Model:
Dry.matter ~ Type + Block
       Df Sum of Sq     RSS     AIC F value  Pr(>F)
<none>               9.9701  5.3660
Type    2    10.946 20.9157 14.7023  5.4892 0.02461 *
Block   5    11.218 21.1884  8.9354  2.2504 0.12881
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> drop1(analysis.ni,test="F")

Model:
Nitrogen ~ Type + Block
       Df Sum of Sq       RSS     AIC F value  Pr(>F)
<none>              0.0083106 -122.25
Type    2 0.0080308 0.0163413 -114.08  4.8317 0.03402 *
Block   5 0.0108023 0.0191128 -117.26  2.5997 0.09315 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> drop1(analysis.gm,test="F")

Model:
Green.matter ~ Type + Block
       Df Sum of Sq    RSS    AIC F value  Pr(>F)
<none>              190.60 58.477
Type    2    260.17 450.77 69.971  6.8251 0.01352 *
Block   5    261.70 452.30 64.032  2.7461 0.08172 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- In all three analyses, the Block effect is insignificant at the 5% test level!
- The impact of Block only materializes when the interdependency between the different yields are taken into account!
- Significance of Type is MUCH stronger in the multivariate model!

Covariance Structures:
Random Effects Models

- For simplicity, we return to a univariate setting.

- **Strength data:**

Three different weigtlifting programs were compared:

  - **RI**: The number of **repetitions** of weightlifting was increased as subjects became stronger.
  - **WI**: The amount of **weight** was increased as subjects became stronger.
  - **CONT**: Control group with no training modifications.

Data were measured at 7 consequtive sessions S1-S7, carried out every other day.

Data from Littel, Freund and Spector (1991).

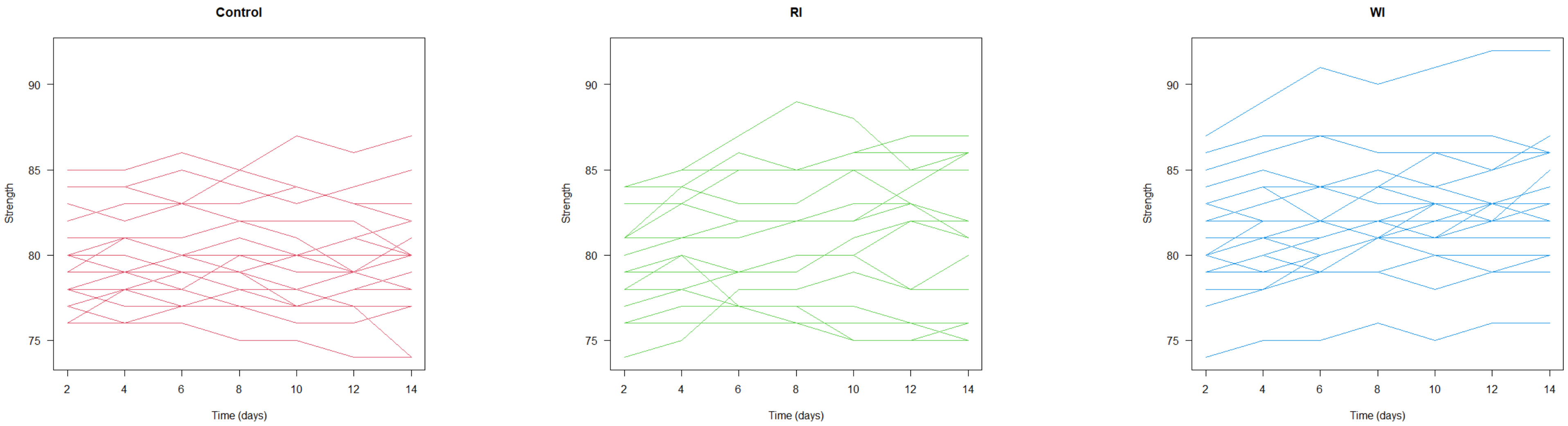
Random Effects Models
Example: Strength data

```
> strength<-read.csv2("Data/Strength data.csv")
> strength$program<-as.factor(strength$program)
> summary(strength)
subject   program         S1              S2              S3              S4       
 Min.   : 1   cont:20   Min.   :74.00   Min.   :75.00   Min.   :75.00   Min.   :75.00  
 1st Qu.:15   ri  :16   1st Qu.:78.00   1st Qu.:79.00   1st Qu.:79.00   1st Qu.:79.00  
 Median :29   wi  :21   Median :80.00   Median :81.00   Median :80.00   Median :81.00  
 Mean   :29             Mean   :80.21   Mean   :80.75   Mean   :80.93   Mean   :81.23  
 3rd Qu.:43             3rd Qu.:83.00   3rd Qu.:83.00   3rd Qu.:83.00   3rd Qu.:84.00  
 Max.   :57             Max.   :87.00   Max.   :89.00   Max.   :91.00   Max.   :90.00  
       S5              S6              S7       
 Min.   :75.00   Min.   :74.00   Min.   :74.00  
 1st Qu.:79.00   1st Qu.:79.00   1st Qu.:79.00  
 Median :81.00   Median :82.00   Median :81.00  
 Mean   :81.25   Mean   :81.18   Mean   :81.32  
 3rd Qu.:84.00   3rd Qu.:83.00   3rd Qu.:84.00  
 Max.   :91.00   Max.   :92.00   Max.   :92.00  
```

- Measurements depend on time, program and subject.

Random Effects Models
Example: Strength data

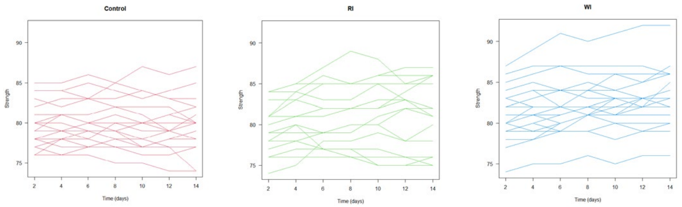
- Measurements depend on time, program and subject.



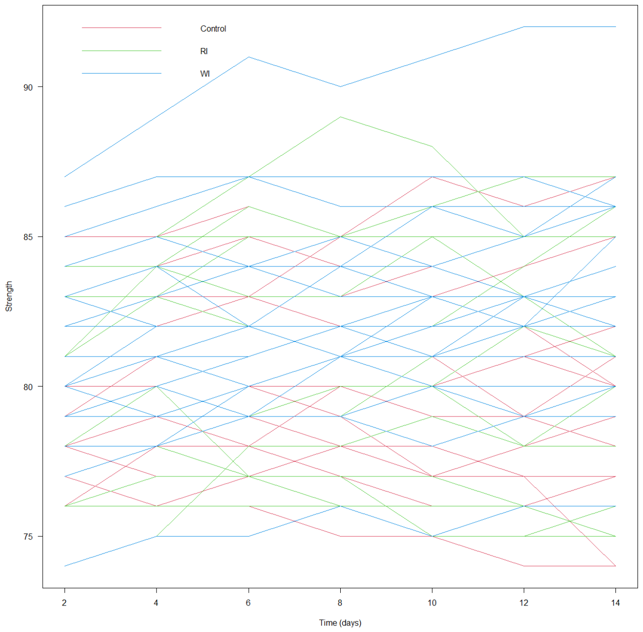
- More or less parallel lines.

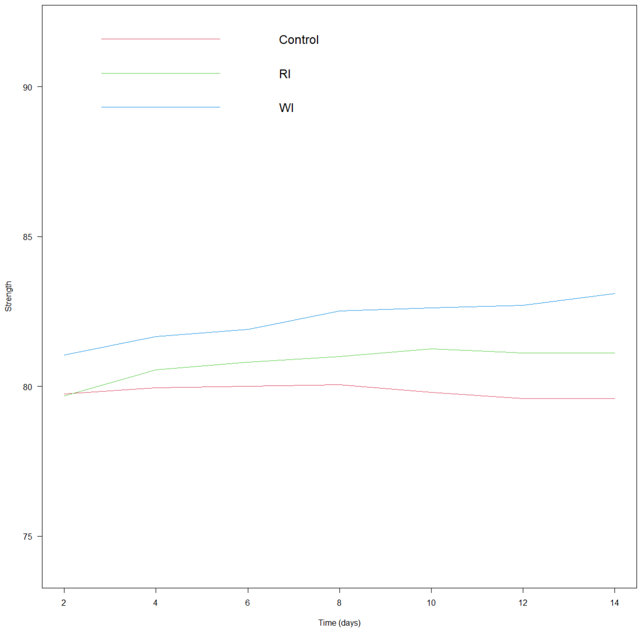
Random Effects Models
Example: Strength data

- Measurements depend on time, program and subject.



- More or less parallel lines means that mean curves make sense!





Random Effects Models
Example: Strength data

- Naiive model:

$$Y_{ij} = \alpha_{program(i)} + \beta \cdot t_j + \gamma \cdot t_j^2 + \delta_{subject(i)} + \varepsilon_{ij}, \varepsilon_{ij} \sim N(0, \sigma^2), i = 1, \dots, 57, j = 1, \dots, 7.$$

- Problem: The model is only valid for these specific subjects!

- We want a prospective model, that investigates the programme effectiveness for future weightlifters.

Random Effects Models
Example: Strength data

- Random effects modeling:

Greek letters

Latin letter

$$Y_{ij} = \alpha_{program(i)} + \beta \cdot t_j + \gamma \cdot t_j^2 + Z_{subject(i)} + \varepsilon_{ij},$$

$$\varepsilon_{ij} \sim N\left(0, \sigma^2\right), i = 1, \dots, 57, j = 1, \dots 7, \qquad Z_i \sim N\left(0, \eta^2\right), i = 1, \dots, 57.$$

- We **RANDOMIZE** the effect of *subject*; We move it from the systematic part of the model to the random part of the model.

- Gain: A prospective model valid for all (also future) subjects;
- Loss: Increased uncertainty.
- Loss: Introction of dependence within subjects.

Random Effects Models
Example: Strength data

$$Y_{ij} = \alpha_{program(i)} + \beta \cdot t_j + \gamma \cdot t_j^2 + Z_i + \varepsilon_{ij},$$
$$\varepsilon_{ij} \sim N(0, \sigma^2), i = 1, \dots, 57, j = 1, \dots 7, \qquad Z_i \sim N(0, \eta^2), i = 1, \dots, 57.$$

Y= mean + between subject variation + within subject variation

$$\begin{aligned} V(Y_{ij}) &= V(mean_{ij} + Z_i + \varepsilon_{ij}) = V(Z_i + \varepsilon_{ij}) = \eta^2 + \sigma^2 \\ Cov(Y_{ij}, Y_{i\ell}) &= Cov(mean_{ij} + Z_i + \varepsilon_{ij}, mean_{i\ell} + Z_i + \varepsilon_{i\ell}) \\ &= Cov(Z_i + \varepsilon_{ij}, Z_i + \varepsilon_{i\ell}) = Cov(Z_i, Z_i) = V(Z_i) = \eta^2 \end{aligned}$$

- Intraclass Correlation Coefficient ICC: How much of the total variation does variation between subjects account for?

$$ICC = \frac{\eta^2}{\eta^2 + \sigma^2}$$

Random Effects Models
Example: Strength data

$$Y_{ij} = \alpha_{program(i)} + \beta \cdot t_j + \gamma \cdot t_j^2 + Z_i + \varepsilon_{ij},$$
$$\varepsilon_{ij} \sim N\left(0, \sigma^2\right), i = 1, \dots, 57, j = 1, \dots 7, \qquad Z_i \sim N\left(0, \eta^2\right), i = 1, \dots, 57.$$

- Within subject variance structure:

$$V\begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{i7} \end{pmatrix} = \sigma^2 I_7 + \eta^2 E_7 = \begin{bmatrix} \sigma^2+\eta^2 & \eta^2 & \cdots & \eta^2 \\ \eta^2 & \ddots & & \vdots \\ \vdots & & \ddots & \eta^2 \\ \eta^2 & \cdots & \eta^2 & \sigma^2+\eta^2 \end{bmatrix}$$

where $E_7 = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}$.

- **Compound symmetry** structure; the variables may be exchanged wíthout changing the structure.

Random Effects Models
Example: Strength data

- Implementation in **R**: Restructuring data

```
> strength2<-data.frame(subject=rep(1:dim(strength)[1],each=7))
> strength2$program<-rep(strength$program,each=7)
> strength2$time<-rep(2*(1:7),dim(strength)[1])
> strength2$strength<-c(t(as.matrix(strength[,3:9])))
> head(strength2)
  subject program time strength
1       1    cont    2       85
2       1    cont    4       85
3       1    cont    6       86
4       1    cont    8       85
5       1    cont   10       87
6       1    cont   12       86
```

- One observation per line

Random Effects Models
Example: Strength data

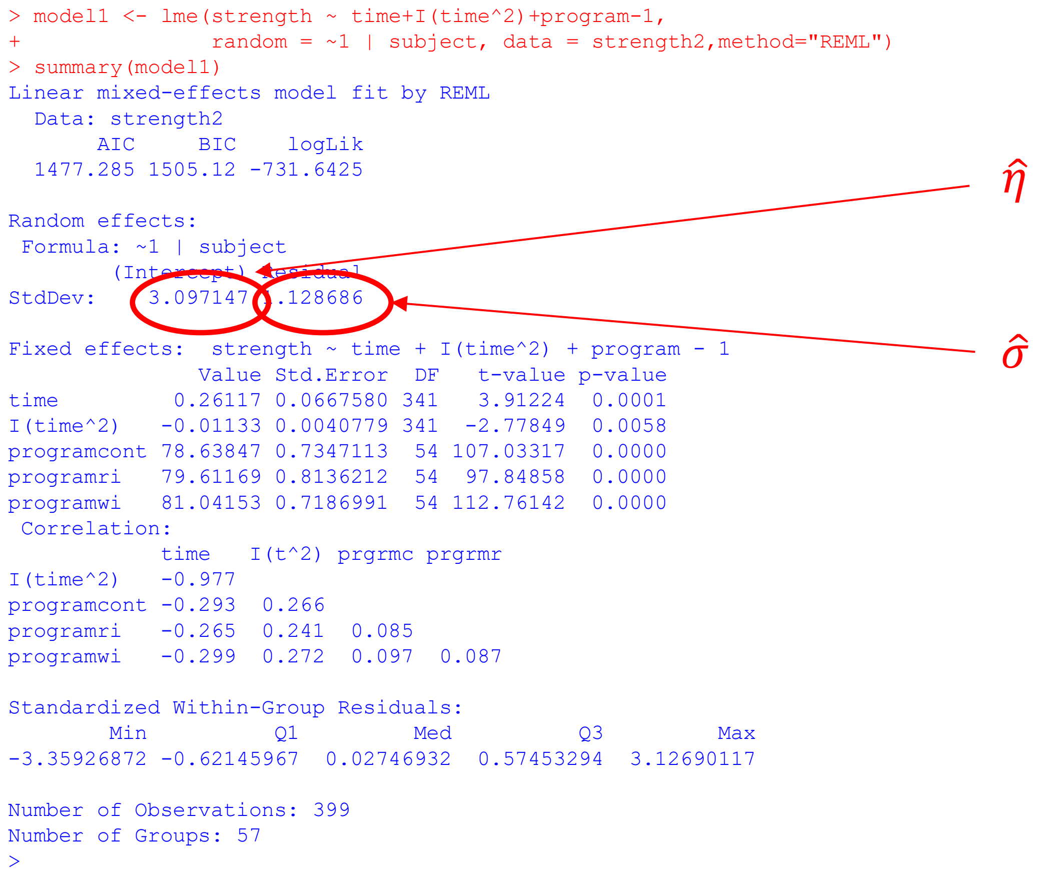
Implementation in **R**: Modeling with the `lme` function

```
> library(nlme)
> model1 <- lme(strength ~ time+I(time^2)+program,
+                random = ~1 | subject, data = strength2,method="ML")

> anova(model1)
            numDF denDF  F-value p-value
(Intercept)     1   340 39860.99  <.0001
time            1   340    31.77  <.0001
I(time^2)       1   340     7.67  0.0059
program         2    54     3.19  0.0488
```

- program is statistically significant, but only barely ($p = 0.05$).

Random Effects Models
Example: Strength data

Implementation in **R**: Post hoc analysis:



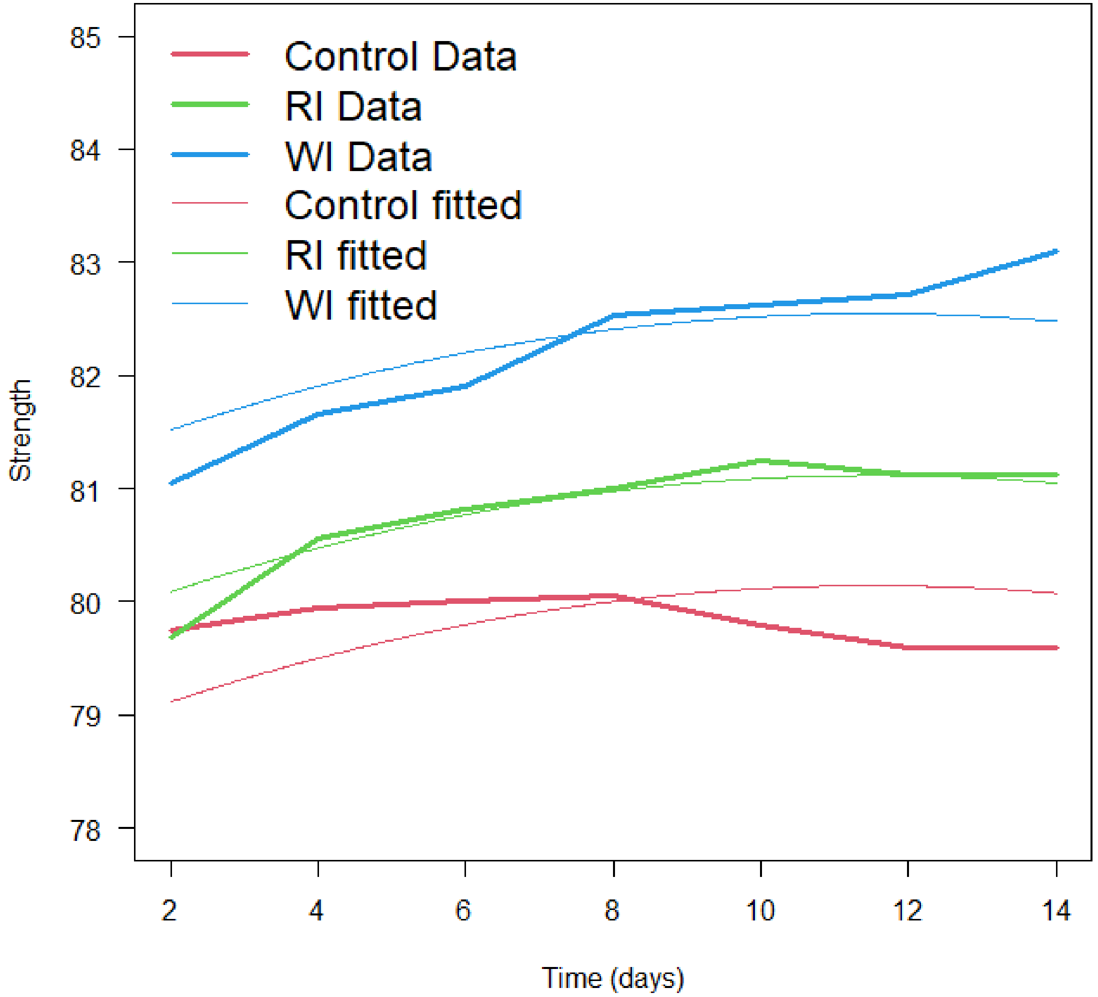
```
> model1 <- lme(strength ~ time+I(time^2)+program-1,
+                random = ~1 | subject, data = strength2,method="REML")
> summary(model1)
Linear mixed-effects model fit by REML
  Data: strength2
       AIC      BIC    logLik
  1477.285 1505.12 -731.6425

Random effects:
 Formula: ~1 | subject
        (Intercept) Residual
StdDev:    3.097147 1.128686

Fixed effects:  strength ~ time + I(time^2) + program - 1
               Value Std.Error  DF   t-value p-value
time         0.26117 0.0667580 341   3.91224  0.0001
I(time^2)   -0.01133 0.0040779 341  -2.77849  0.0058
programcont 78.63847 0.7347113  54 107.03317  0.0000
programri   79.61169 0.8136212  54  97.84858  0.0000
programwi   81.04153 0.7186991  54 112.76142  0.0000
 Correlation:
            time   I(t^2) prgrmc prgrmr
I(time^2)   -0.977
programcont -0.293  0.266
programri   -0.265  0.241  0.085
programwi   -0.299  0.272  0.097  0.087

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-3.35926872 -0.62145967  0.02746932  0.57453294  3.12690117

Number of Observations: 399
Number of Groups: 57
>
```

$$ICC = \frac{3.097147^2}{3.097147^2 + 1.128686^2} = 0.88$$

Random Effects Models
Example: Strength data

Implementation in R: Post hoc analysis:

```
> summary(model1)$tTable
 Value    Std.Error  DF     t-value      p-value
time          0.26117377 0.066758032 341   3.912245 1.103833e-04
I(time^2)   -0.01133041 0.004077896 341  -2.778494 5.763578e-03
programcont 78.63847118 0.734711251  54 107.033166 1.446154e-64
programri   79.61168546 0.813621243  54  97.848583 1.793276e-62
programwi   81.04153240 0.718699138  54 112.761416 8.767784e-66
> (my.coef<-summary(model1)$tTable[,1])
       time   I(time^2) programcont   programri   programwi
 0.26117377 -0.01133041 78.63847118 79.61168546 81.04153240
```

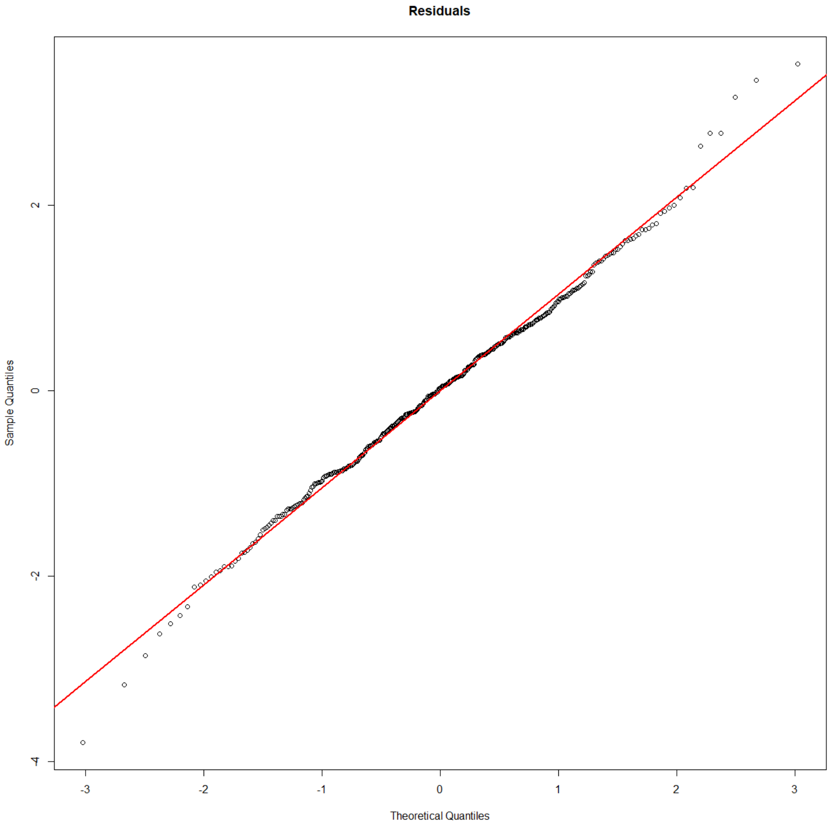


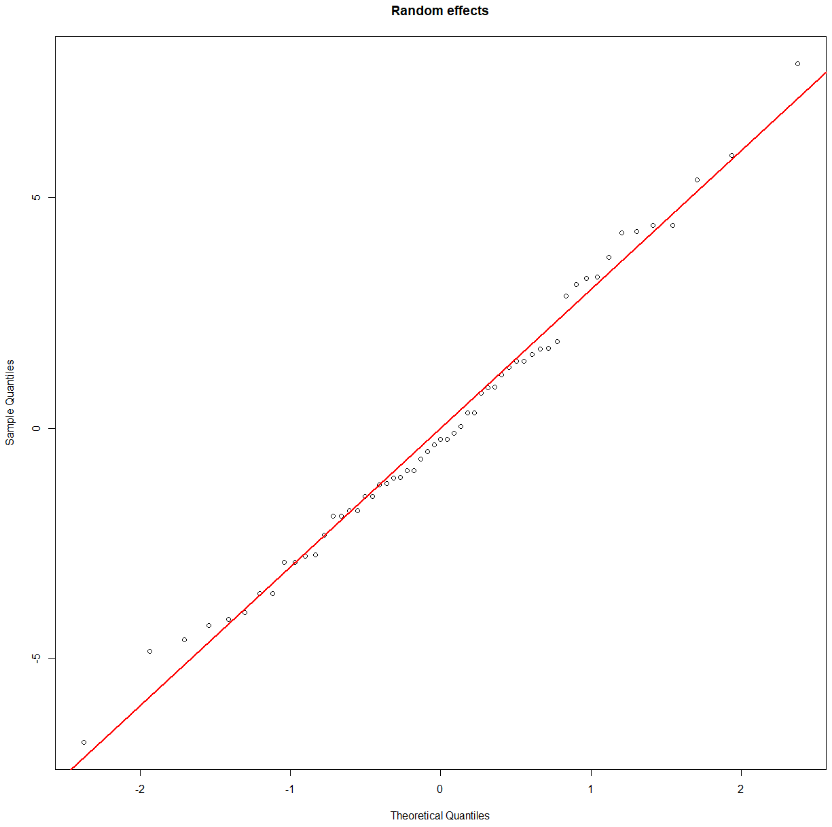
Random Effects Models
Example: Strength data

Implementation in **R**: Model control.

Residuals ($\hat{\varepsilon}$): `residuals(model1)`.

Estimated random effects ($\hat{Z}$): `ranef(model1)`.





Random Effects Models
Example: Strength data

- Within subject variance structure:

$$V\begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{i7} \end{pmatrix} = \sigma^2 I_7 + \eta^2 E_7 = \begin{bmatrix} \sigma^2+\eta^2 & \eta^2 & \cdots & \eta^2 \\ \eta^2 & \ddots & & \vdots \\ \vdots & & \ddots & \eta^2 \\ \eta^2 & \cdots & \eta^2 & \sigma^2+\eta^2 \end{bmatrix}$$

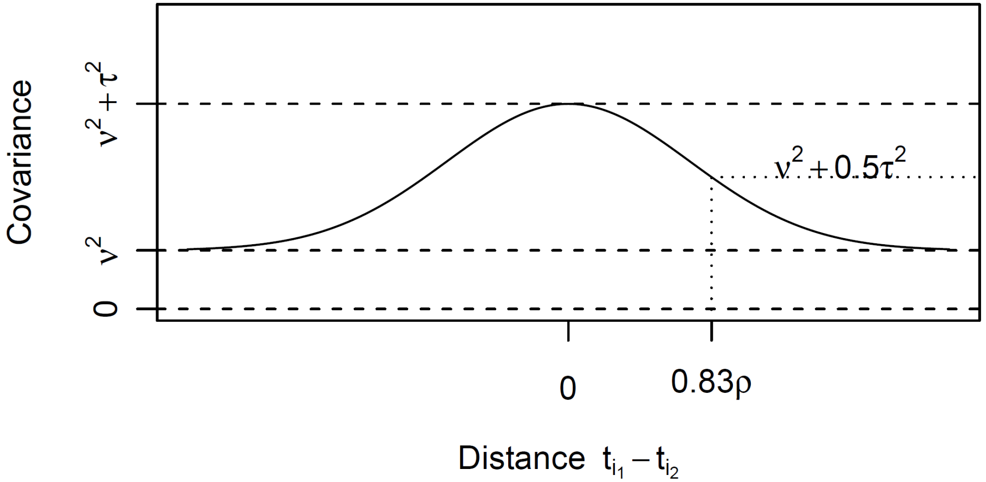
Is it realistic that the covariance is the same for all time points?

- Covariance may decline with time.
- Trick: Model the covariance directly, replacing E_7 with kernel entries that models decline in time.

Random Effects Models: Advanced Correlation Structures

- **Spatial covariance structures**, depending on how far away in time the variables are from each other.
- We adopt the structure from the strength2 dataset.
- Gaussian spatial covariance structure:

$$Cov(Y_{i_1}, Y_{i_2}) = \begin{cases} 0 & if\ subject(i_1) \neq subject(i_2) \\ \nu^2 + \tau^2 exp\left(-\frac{\left(t_{i_1} - t_{i_2}\right)^2}{\rho^2}\right) & if\ subject(i_1) = subject(i_2), i_1 \neq i_2 \\ \nu^2 + \tau^2 + \sigma^2 & if\ i_1 = i_2 \end{cases}$$



Random Effects Models
Example: Strength data

- Model for the entire data vector:

$$Y \sim N_{399}(\mu, V)$$

where

$$\mu_i = \alpha_{program(i)} + \beta t_i + \gamma t_i^2,$$

and

$$V_{i,j} = \begin{cases} 0 & if\ subject(i) \neq subject(j) \\ \nu^2 + \tau^2 exp\left(-\frac{\left(t_i - t_j\right)^2}{\rho^2}\right) & if\ subject(i) = subject(j), i \neq j \\ \nu^2 + \tau^2 + \sigma^2 & if\ i = j \end{cases}$$

Random Effects Models
Example: Strength data

- Implementation in **R**:

```
> model2<-lme(strength ~ time+I(time^2)+program,random=~1|subject,
+            correlation=corGaus(form=~time|subject,nugget=T),method="ML",
+            data=strength2)

> anova(model2)
            numDF denDF  F-value p-value
(Intercept)     1   340 40750.61  <.0001
time            1   340    13.56  0.0003
I(time^2)       1   340     7.28  0.0073
program         2    54     3.34  0.0429
```

- Stronger significance of program ($p = 0.04$).

Random Effects Models
Example: Strength data

- Parameter estimates:

```
>summary(model2)
Linear mixed-effects model fit by maximum likelihood
  Data: strength2
       AIC      BIC    logLik
  1293.076 1328.976 -637.5379

Random effects:
 Formula: ~1 | subject
        (Intercept) Residual
StdDev:    2.721507$(= \hat{\nu})$ 1.712461$\left(= \sqrt{\hat{\sigma}^2 + \hat{\tau}^2}\right)$

Correlation Structure: Gaussian spatial correlation
 Formula: ~time | subject
 Parameter estimate(s):
   range          nugget
8.5026559 $(= \hat{\rho}^2)$ 0.1394383 $\left(= \frac{\hat{\sigma}^2}{\hat{\sigma}^2+\hat{\tau}^2}\right)$
Fixed effects:  strength ~ time + I(time^2) + program
…
```

$$\hat{\boldsymbol{\nu}}^2 = 7.4066;\ \hat{\boldsymbol{\tau}}^2 = 2.523617;\ \hat{\boldsymbol{\sigma}}^2 = 0.408906$$

Total variance:

$$\hat{\boldsymbol{\nu}}^2 + \hat{\boldsymbol{\tau}}^2 + \hat{\boldsymbol{\sigma}}^2 = 10.34$$

Random Effects Models: Advanced Correlation Structures

- R has a number of built-in correlation structures:

Name	Covariance term
Gaussian	$\nu^2 + \tau^2 exp\left(-\frac{(t_i - t_j)^2}{\rho^2}\right)$
Exponential	$\nu^2 + \tau^2 exp\left(-\frac{\lvert t_i - t_j \rvert}{\rho^2}\right)$
AR(1)	$\nu^2 + \rho^{\lvert t_i - t_j \rvert}$
Unstructured	$\tau_{i,j}$

Random Effects Models: Advanced Correlation Structures

- **Which correlation structure to choose?** – stationarity.
- For all the listed covariance structures, the value of $Cov(Y_t, Y_t + u)$ does not depend on the time t; only on the time difference u.
- We say that the error proces ε_t is *weakly stationary of order 2*;
- *weakly* because the stationarity is defined from moments, and not distributions (for Gaussian processes this is the same though);
- *of order 2* because the stationarity is defined through 2nd order moments (variance, covariance).

Random Effects Models: Advanced Correlation Structures The Semivariogram

- **Which correlation structure to choose?** – The semivariogram plots $\gamma(u) = \frac{1}{2}V(Y_t - Y_{t+u}), u > 0.$

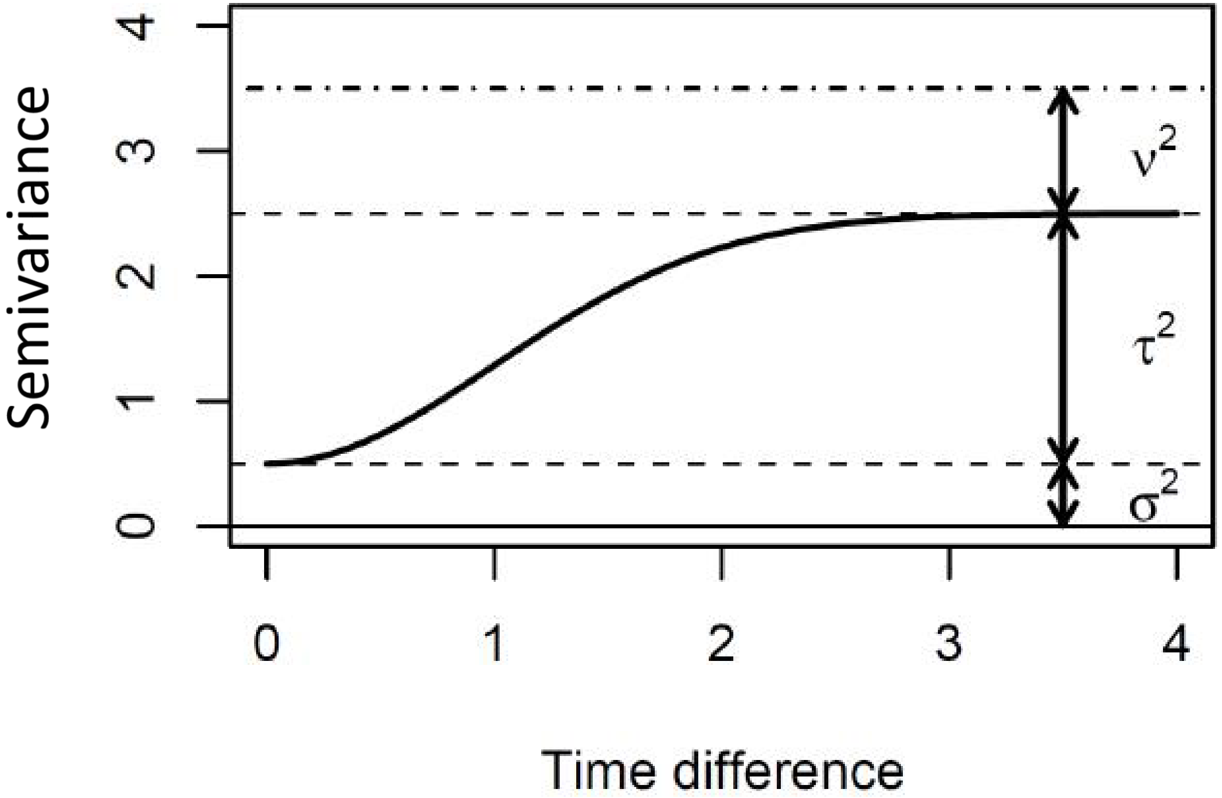
$$\begin{aligned}
\gamma(u) &= \frac{1}{2}V(Y_t - Y_{t+u}) \\
&= \frac{1}{2}\big(V(Y_t) + V(Y_{t+u}) - 2Cov(Y_t, Y_{t+u})\big) \\
&= V(Y_t) - Cov(Y_t, Y_{t+u}) \\
&= \nu^2 + \tau^2 + \sigma^2 - \nu^2 - \tau^2\lambda(u) \\
&= \sigma^2 + \tau^2\big(1 - \lambda(u)\big)
\end{aligned}$$

where $\lambda(u) = exp\big(-u^2/\rho^2\big),\ exp(-u/\rho),\ \rho^u$ respectively, for Gaussian, exponential and AR(1) spatial decay of correlations.

- Note that for the **compound symmetry model**, $\gamma(u) = \sigma^2$, independent of u.

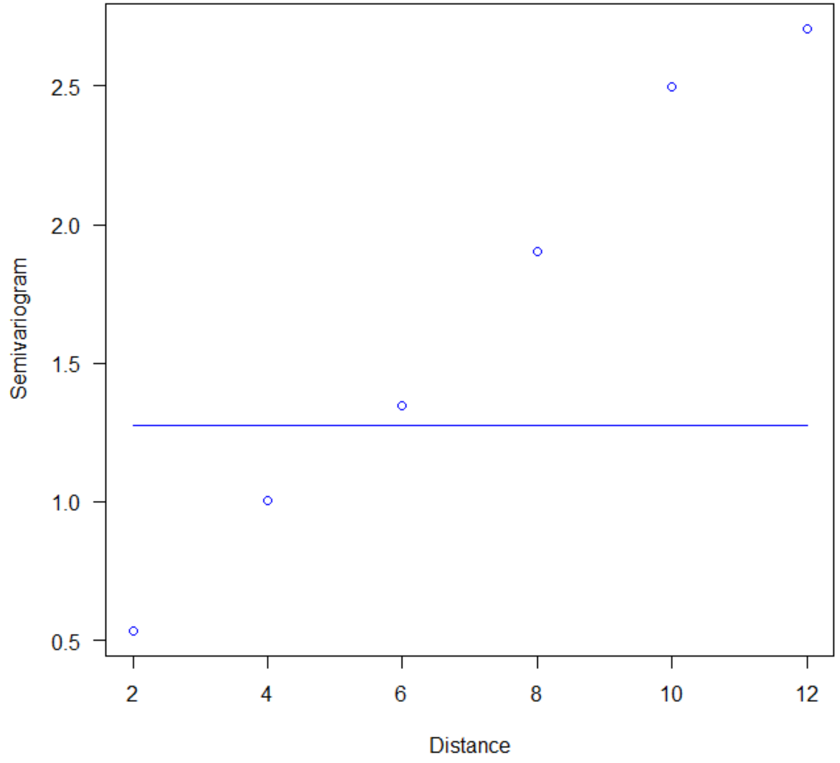
Random Effects Models: Advanced Correlation Structures The Semivariogram

- Theoretical semivariogram with Gaussian correlation, $\lambda(u) = exp\left(-u^2/\rho^2\right)$:



Random Effects Models: Advanced Correlation Structures The Semivariogram

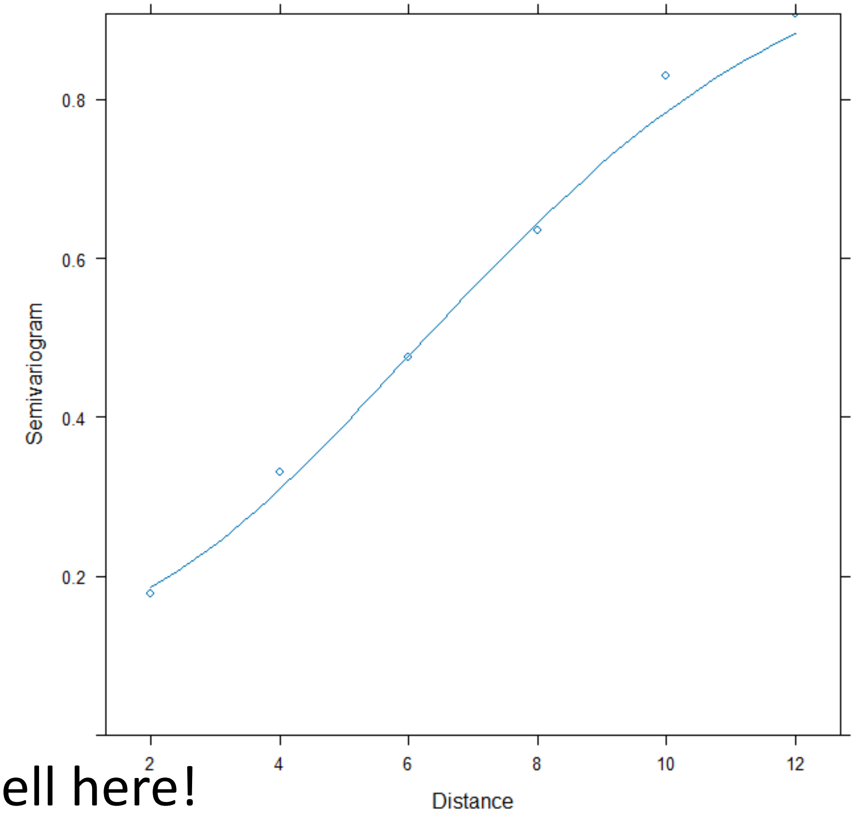
- Empirical variogram with compound symmetry comparison:



- Not good; the theoretical line doesn’t fit the empirical points.

Random Effects Models: Advanced Correlation Structures The Semivariogram

- Semivariogram with Gaussian decay:



- Theory and practice fits well here!
- We select the model with a Spatial Gaussian Covariance structure.

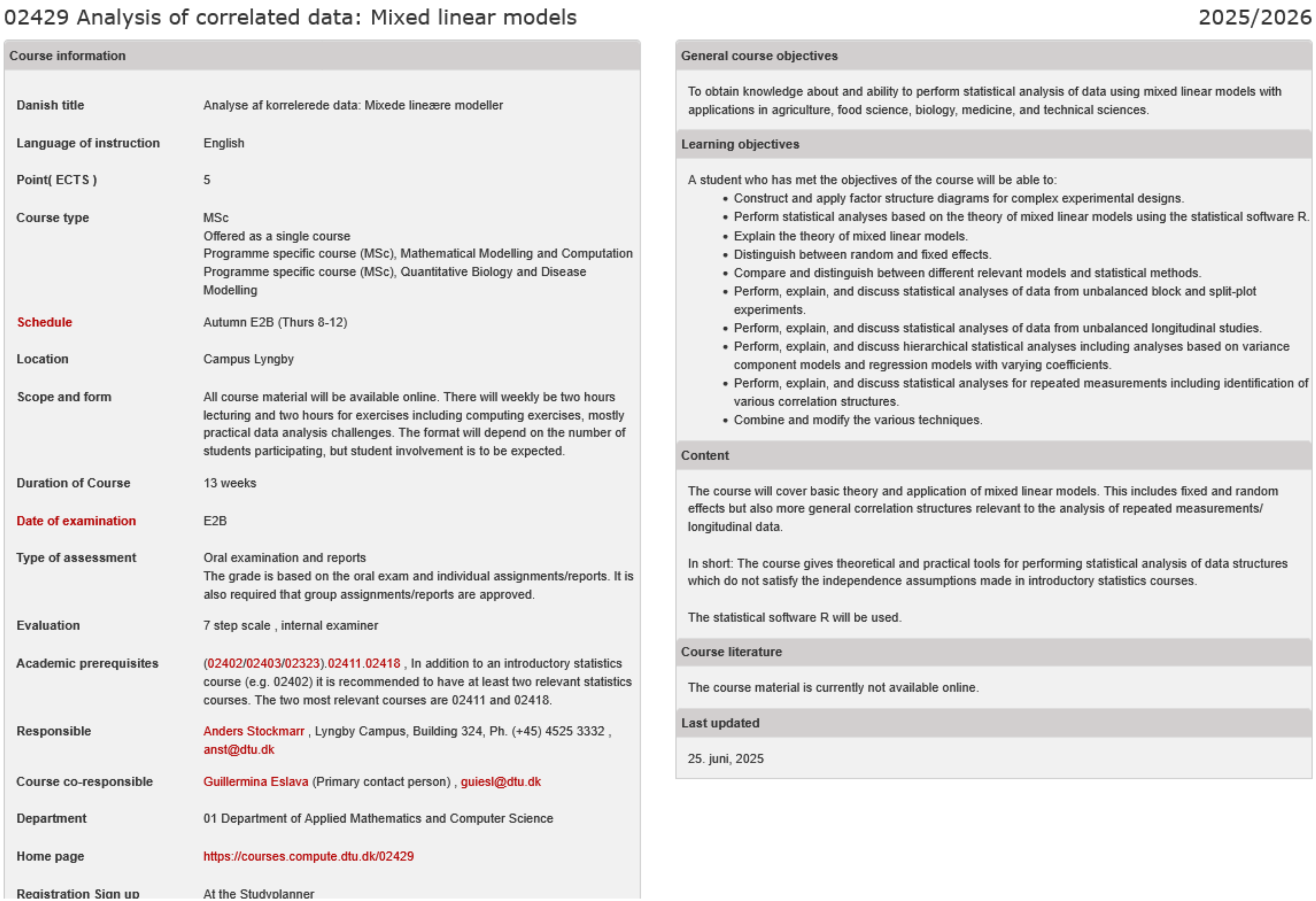
Random Effects Models: Advanced Correlation Structures

- Pro’s and Cons of approaches:

- Compound symmetry model:
- **Pro:**
  - Uses all the data
  - Easily interpretable
  - Often good for short time series.
- **Con:**
  - Often not good for long time series.

- Spatial correlation structure model:
- **Pro:**
  - Uses all the data
  - Works for long and short time series.
- **Con:**
  - Requires appropriate choice of Covariance structure

Random Effects Models at DTU



02429 Analysis of correlated data: Mixed linear models 2025/2026

Course information

Danish title	Analyse af korrelerede data: Mixede lineære modeller
Language of instruction	English
Point(ECTS)	5
Course type	MSc Offered as a single course Programme specific course (MSc), Mathematical Modelling and Computation Programme specific course (MSc), Quantitative Biology and Disease Modelling
Schedule	Autumn E2B (Thurs 8-12)
Location	Campus Lyngby
Scope and form	All course material will be available online. There will weekly be two hours lecturing and two hours for exercises including computing exercises, mostly practical data analysis challenges. The format will depend on the number of students participating, but student involvement is to be expected.
Duration of Course	13 weeks
Date of examination	E2B
Type of assessment	Oral examination and reports The grade is based on the oral exam and individual assignments/reports. It is also required that group assignments/reports are approved.
Evaluation	7 step scale , internal examiner
Academic prerequisites	(02402/02403/02323).02411.02418 , In addition to an introductory statistics course (e.g. 02402) it is recommended to have at least two relevant statistics courses. The two most relevant courses are 02411 and 02418.
Responsible	Anders Stockmarr , Lyngby Campus, Building 324, Ph. (+45) 4525 3332 , anst@dtu.dk
Course co-responsible	Guillermina Eslava (Primary contact person) , guiesl@dtu.dk
Department	01 Department of Applied Mathematics and Computer Science
Home page	https://courses.compute.dtu.dk/02429
Registration Sign up	At the Studyplanner

General course objectives

To obtain knowledge about and ability to perform statistical analysis of data using mixed linear models with applications in agriculture, food science, biology, medicine, and technical sciences.

Learning objectives

A student who has met the objectives of the course will be able to:

- Construct and apply factor structure diagrams for complex experimental designs.
- Perform statistical analyses based on the theory of mixed linear models using the statistical software R.
- Explain the theory of mixed linear models.
- Distinguish between random and fixed effects.
- Compare and distinguish between different relevant models and statistical methods.
- Perform, explain, and discuss statistical analyses of data from unbalanced block and split-plot experiments.
- Perform, explain, and discuss statistical analyses of data from unbalanced longitudinal studies.
- Perform, explain, and discuss hierarchical statistical analyses including analyses based on variance component models and regression models with varying coefficients.
- Perform, explain, and discuss statistical analyses for repeated measurements including identification of various correlation structures.
- Combine and modify the various techniques.

Content

The course will cover basic theory and application of mixed linear models. This includes fixed and random effects but also more general correlation structures relevant to the analysis of repeated measurements/ longitudinal data.

In short: The course gives theoretical and practical tools for performing statistical analysis of data structures which do not satisfy the independence assumptions made in introductory statistics courses.

The statistical software R will be used.

Course literature

The course material is currently not available online.

Last updated

25. juni, 2025

Exercises

- Exam 2023, questions 1-4.