Lecture 2: Random Effects and Hierarchical Structures

Dankmar Böhning

Southampton Statistical Sciences Research Institute
University of Southampton, UK

S^3RI, 12-13 June 2014

Crossed Factors

- two factors A (a levels) and B (b levels)
- **Example:**
- experiment is done to study effect of temperature on yield of tomato plants
- A room temperature, B soil temperature
- both have 2 levels: **high** and **low**

Crossed Factors

Definition

experiment has factors crossed if all combinations of factors are available

Example

in the example with soil and room temperature:

$$(high, high), (high, low), (low, high), (low, low)$$

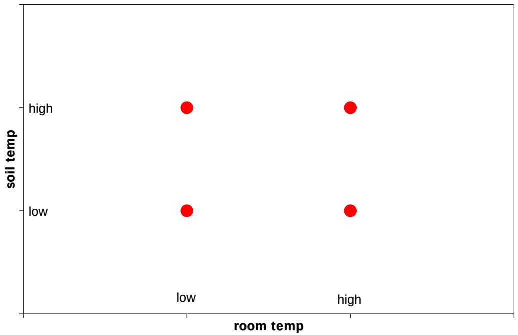


Figure: Example of two factor experiment with both factors crossed

Nested Factors

- two factors A (a levels) and B (b levels),
- but B is nested within A
- **Example:**
- a company operates two machines and 4 operators work with these machines
- **but:** only the first two operators (1 and 2) work on machine 1,
- the second two operators (3 and 4) on machine 2
- company is interested in the effect of
- A machine and B operator on machine product
- **important:** operator is *nested* within machine

Nested Factors

Definition

experiment has factor b nested within A nested if level of B varies only within A

Example

in the example with machine and operator:

$$(o1, m1), (o2, m1), (o3, m2), (o4, m2)$$

where m indicates machine and o operator

Nested Factors

Definition

experiment has factor b nested within A nested if level of B varies only within A

Example

in the example with machine and operator:

$$(o1, m1), (o2, m1), (o3, m2), (o4, m2)$$

where m indicates machine and o operator

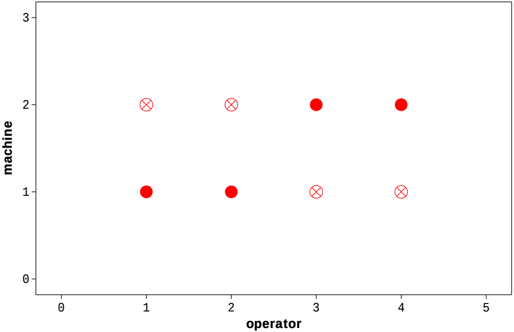


Figure: Example of two factor experiment with factor B operator nested within A machine

Nested Factors

- two factors D (for doctor) and P (for patient),
- but P is nested within D
- since not every doctor consults every patient in the hospital
- **important:** patient is *nested* within doctor
-
- two factors W (for ward) and P (for patient),
- but P is nested within W
- since patients stay within their wards in the hospital
- **important:** patient is *nested* within ward

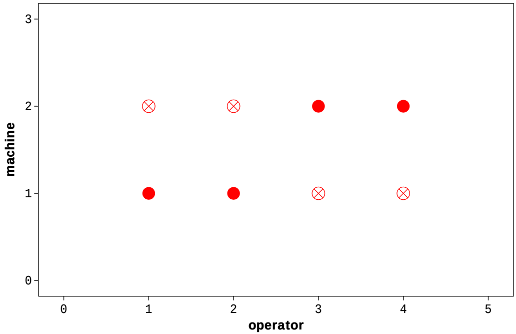


Figure: Example of two factor experiment with factor B operator nested within A machine

ANOVA-Model for Crossed Factors

Model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

with $\epsilon_{ijk} \sim N(0, \sigma^2)$ and usual constraints on main effects α_i, β_j and interactions $(\alpha\beta)_{ij}$ (if they are fixed)

- $i = 1, ..., a$
- $j = 1, ..., b$
- $k = 1, ..., n$

so that there are $n(n_j)$ observations per factor (j) combination

ANOVA-Model for Nested Factors

Model

$$Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \epsilon_{ijk}$$

with $\epsilon_{ijk} \sim N(0, \sigma^2)$ and usual constraints on main effects α_i, $\beta_{j(i)}$ (if they are fixed), Factor B is nested within A

- $i = 1, ..., a$
- $j = 1, ..., b$
- $k = 1, ..., n$ or $k = 1, ..., n_j$

so that there are $n(n_j)$ observations per factor (j) combination

- note that if B is nested A one can test for a main effect for A but **not for an interaction with B separately from the main effect of B**
- this is because B is changing only within A

Fixed or Random Effects?

when should we consider a factor random and when fixed?

no absolute rules exist. However, it is beneficial to consider a factor as **random** if

- the levels of the factor can be considered a sample from a much larger population
- the levels of the factor increase with the sample size

it is appropriate to consider a factor as **fixed** if

- there is specific interest in the levels of the factor
- the levels of the factor (intervention, therapy) remain fixed when the sample size increases

An example

- Data were collected on patients suffering from rheumatoid arthritis (RA) of the hands
- Computer program used to assess degree of severity
- Program analyses x-rays and determines mean joint space
- Only the left hands of patients were analysed.

the data have the following hierarchical structure:

- 4 patients randomly selected
- for each patient 3 x-rays were taken
- each x-ray was analysed by the computer twice

Data:

Patient	X-ray	computer–analysis	
1	1	0.626	0.625
1	2	0.846	0.894
1	3	0.982	0.949
2	1	0.867	0.928
2	2	0.976	0.930
2	3	0.986	0.994
3	1	1.168	1.209
3	2	1.313	1.324
3	3	1.214	1.406
4	1	0.198	0.173
4	2	0.234	0.203
4	3	0.179	0.177

Model

for the computer analysis Y_{ijk} for patient i, x-ray j, and computer analysis k:

$$Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \epsilon_{ijk}$$

with

- a patient random effect $\alpha_i \sim N(0, \sigma_P^2)$
- an x-ray random effect $\beta_{i(j)} \sim N(0, \sigma_X^2)$ nested in the patient effect
- with a random error $\epsilon_{ijk} \sim N(0, \sigma^2)$

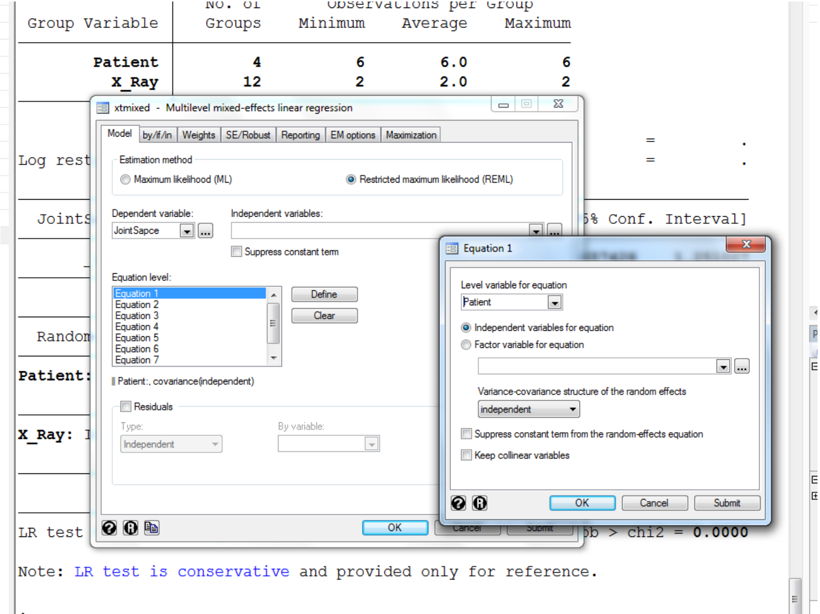
Model

$$Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \epsilon_{ijk}$$

since α_i and $\beta_{j(i)}$ assumed independent

- $$Var(Y_{ijk}) = \sigma_P^2 + \sigma_X^2 + \sigma^2$$

hence we have a variance component model with **three variance components**


No. of
Observations per Group
Group Variable
Groups
Minimum
Average
Maximum
Patient
4
6
6.0
6
X_Ray
12
2
2.0
2
Log rest
Joints
Random
Patient:
X_Ray: I
LR test
5% Conf. Interval]
ob > chi2 = 0.0000
Note: LR test is conservative and provided only for reference.
xtmixed - Multilevel mixed-effects linear regression
Model
by/if/in
Weights
SE/Robust
Reporting
EM options
Maximization
Estimation method
Maximum likelihood (ML)
Restricted maximum likelihood (REML)
Dependent variable:
Independent variables:
JointSapce
Suppress constant term
Equation level:
Equation 1
Equation 2
Equation 3
Equation 4
Equation 5
Equation 6
Equation 7
Define
Clear
|| Patient:, covariance(independent)
Residuals
Type:
Independent
By variable:
OK
Cancel
Submit
Equation 1
Level variable for equation
Patient
Independent variables for equation
Factor variable for equation
Variance-covariance structure of the random effects
independent
Suppress constant term from the random-effects equation
Keep collinear variables
OK
Cancel
Submit

```
Mixed-effects REML regression                  Number of obs      =        24
```

Group Variable	No. of Groups	Observations per Group Minimum	Average	Maximum
Patient	4	6	6.0	6
X_Ray	12	2	2.0	2

```
                                               Wald chi2(0)       =         .
Log restricted-likelihood =  18.352221         Prob > chi2        =         .
```

JointSapce	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_cons	.808375	.2258369	3.58	0.000	.3657428	1.251007

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
Patient: Identity				
var(_cons)	.2007867	.1665807	.0394952	1.020765
X_Ray: Identity				
var(_cons)	.0086354	.0048522	.0028708	.0259759
var(Residual)	.0020646	.0008429	.0009276	.0045956

```
LR test vs. linear regression:      chi2(2) =    64.05   Prob > chi2 = 0.0000
```

Fitting various variance component models:

Model	σ_X^2	σ_P^2	σ^2	$\log L$	LRT
P,X	0.0086	0.2008	0.0021	18.3522	64.05§
P	-	0.2025	0.0090	12.6075	11.49
-	-	-	0.1675	-13.6741	52.55

§ null hypothesis is here that all random effect variance components are zero

Mixed Model

$$Y = X\beta + A\alpha + \epsilon$$

- Y vector of responses
- ϵ vector of mean-zero normal errors
- X design matrix of fixed effects
- β vector of fixed effect parameters
- A design matrix of random effects
- α vector of mean zero normal random effect parameters

amount of purity data:

	Determination of impurity (in%)			
Batch	1	2	3	4
1	3.28	3.09	3.03	3.07
2	3.52	3.48	3.38	3.43
3	2.91	2.80	2.76	2.85
4	3.34	3.38	3.23	3.31
5	3.28	3.14	3.25	3.21
6	2.98	3.01	3.13	2.95

illustration

$$\begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{14} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{24} \\ \dots \\ Y_{61} \\ Y_{62} \\ Y_{63} \\ Y_{64} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ \dots \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \mu + \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ \dots \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \\ \alpha_6 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{14} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \\ \epsilon_{24} \\ \dots \\ \epsilon_{61} \\ \epsilon_{62} \\ \epsilon_{63} \\ \epsilon_{64} \end{pmatrix}$$

Mixed Model

$$Y = X\beta + A\alpha + \epsilon$$

Note that

$$E(Y) = X\beta$$

let

- covariance matrix of ϵ be R
- covariance matrix of α be G

then the covariance matrix V of Y is given as

$$V = AGA' + R$$

estimation:

estimation in linear mixed models is based upon the **multivariate normal log-likelihood**

$$-2\log L = constant + log|V| + (Y - X\beta)'V^{-1}(Y - X\beta)$$

another method is preferred when modeling covariance structures such as variance component models called **restricted maximum likelihood estimation (REML)** which uses a correction factor in the multivariate normal likelihood:

$$-2\log RL = constant + \log|\mathbf{X'V^{-1}X}| + \log|V| + (Y - X\beta)'V^{-1}(Y - X\beta)$$

estimation:

- REML corrects for estimating the mean structures
- gives unbiased estimates of variance components in classical balanced settings
- where MLEs are not unbiased
- both methods give consistent estimators
- for modelling covariance structures we use REML
- for modelling mean structures we us MLE

model evaluation

- for model assessment we will use criteria that compromise between **model fit** and **model complexity**
- Akaike information criterion

$$AIC = -2 \log L + 2k$$

- Bayesian Information criterion

$$BIC = -2 \log L + k \log n$$

- where k is the number of parameters in the model
- and n is the number of clustered observations
- we seek a model for which AIC and/or BIC are small

model assessment:

Model	σ_X^2	σ_P^2	σ^2	$\log L$	LRT	AIC	BIC
P,X	0.0086	0.2008	0.0021	18.35	64.05	-28.70	-29.54
P	-	0.2025	0.0090	12.61	11.49	-19.22	-19.84
-	-	-	0.1675	-13.67	52.55	31.35	30.93