Lecture 3: Random Effects and Hierarchical Structures

Dankmar Böhning

Southampton Statistical Sciences Research Institute
University of Southampton, UK

S^{3}RI, 11 - 12 December 2014

Review: data with simple random effects structure

Crossed and Nested Factors

GLM-Model for Crossed Factors

An Example

data with simple random effects structure

consider the following study data:

- interest is in the amount of **impurity** in a pharmaceutical product
- data arise in form of batches of material as they come off the production line
- 6 batches are randomly selected
- 4 determinations are made per batch

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

questions of interest

- to determine the **average amount of impurity**
- **batch effect**?
- how large is **variation between batches** ?

ONEWAY fixed effect model

$$Y_{ij} = \mu + \beta_i + \epsilon_{ij}$$

- $i = 1, \cdots, 6$, $j = 1, \cdots, 4$
- β_i unknown fixed parameters, $\sum_i \beta_i = 0$
- random error $\epsilon_{ij} \sim N(0, \sigma^2)$
-

$$E(Y_{ij}) = \mu + \beta_i$$

Problems with the ONEWAY fixed effect model

- number of parameters increases with the number of batches
- interest is **not** in a specific effect but more in a general batch effect
- model assumes **independence** of observations within batches
- variance of observations is determined by variance of errors

$$Var(Y_{ij}) = Var(\epsilon_{ij}) = \sigma^2$$

and might likely underestimate variance

- hence confidence intervals for average impurity amount might be too small

more suitable is the ONEWAY random effects model

$$Y_{ij} = \mu + \beta_i + \epsilon_{ij}$$

- $i = 1, \cdots, 6$, $j = 1, \cdots, 4$
- $\beta_i \sim N(0, \sigma_B^2)$ are **random effects**
- random error $\epsilon_{ij} \sim N(0, \sigma^2)$
- α_i and ϵ_{ij} independent
-

$$E(Y_{ij}) = \mu$$

ONEWAY random effects model

$$Y_{ij} = \mu + \beta_i + \epsilon_{ij}$$

- $Var(Y_{ij}) = Var(\beta_i) + Var(\epsilon_{ij}) = \sigma_B^2 + \sigma^2$
- model is a **variance components model**
- covariance **between batches** is 0

$$cov(Y_{ij}, Y_{\ell k}) = 0$$

if $\ell \neq i$, $j \neq k$

ONEWAY random effects model

$$Y_{ij} = \mu + \beta_i + \epsilon_{ij}$$

- covariance **within batches is not 0** $(j \neq k)$:

$$cov(Y_{ij}, Y_{ik}) = E(\alpha_i^2) + E(\alpha_i \epsilon_{ij}) E(\alpha_i \epsilon_{ik}) + E(\epsilon_{ik} \epsilon_{ij}) = \sigma_B^2$$

- hence random effects model is suitable to model within batches correlation (autocorrelation model)

ONEWAY random effects model

random effects

$$Y_{ij} = \mu + \underbrace{\beta_i}_{\text{random}} + \epsilon_{ij}$$

fixed effects

$$Y_{ij} = \mu + \underbrace{\beta_i}_{\text{fixed}} + \epsilon_{ij}$$

- fixed effects models has as many parameters β_i as there are levels of the factor
- potentially many parameters
- random effects model has only **one parameter** σ_B^2

ONEWAY vs. GLM with random effects

random effects with normal errors

$$Y_{ij} = \mu + \underbrace{\beta_i}_{\text{random}} + \epsilon_{ij}$$

random effects with Poisson errors

$$\log E(Y_{ij})_{|\beta_i} = \mu + \underbrace{\beta_i}_{\text{random}}$$

- in both cases we assume $\beta_i \sim N(0, \sigma_B^2)$
- **but** in the GLM case we have no longer the additive normal random error
- and the variance decomposition $Var(Y_{ij}) = \sigma^2 + \sigma_B^2$ **no longer holds**

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:

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

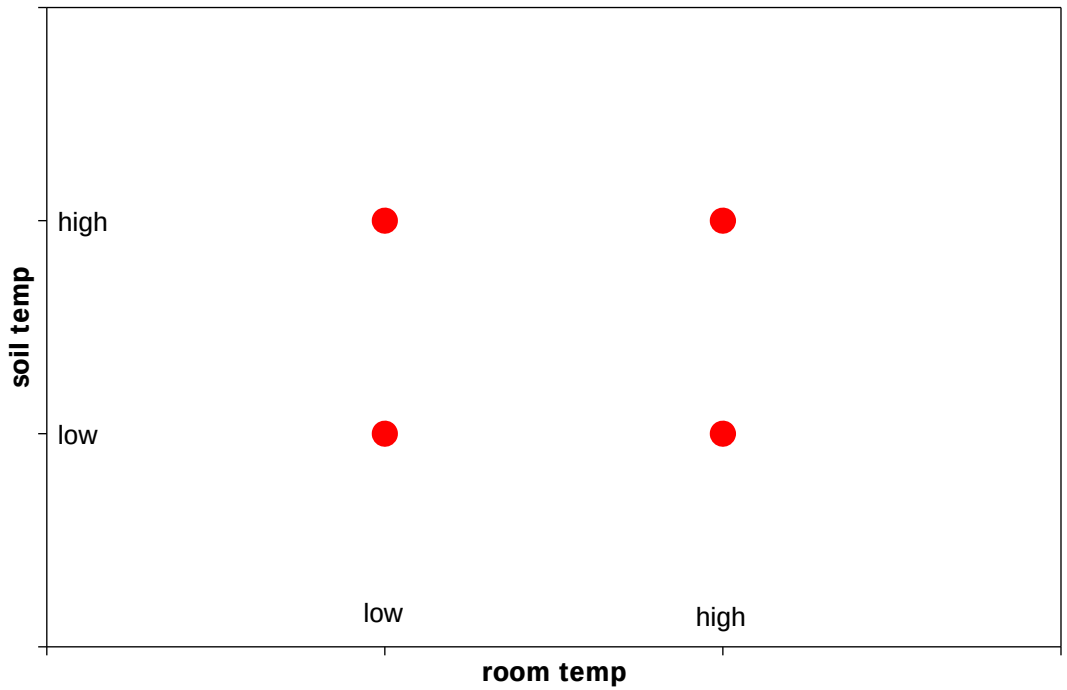


Figure: Example of two factor experiment with both factors crossed

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

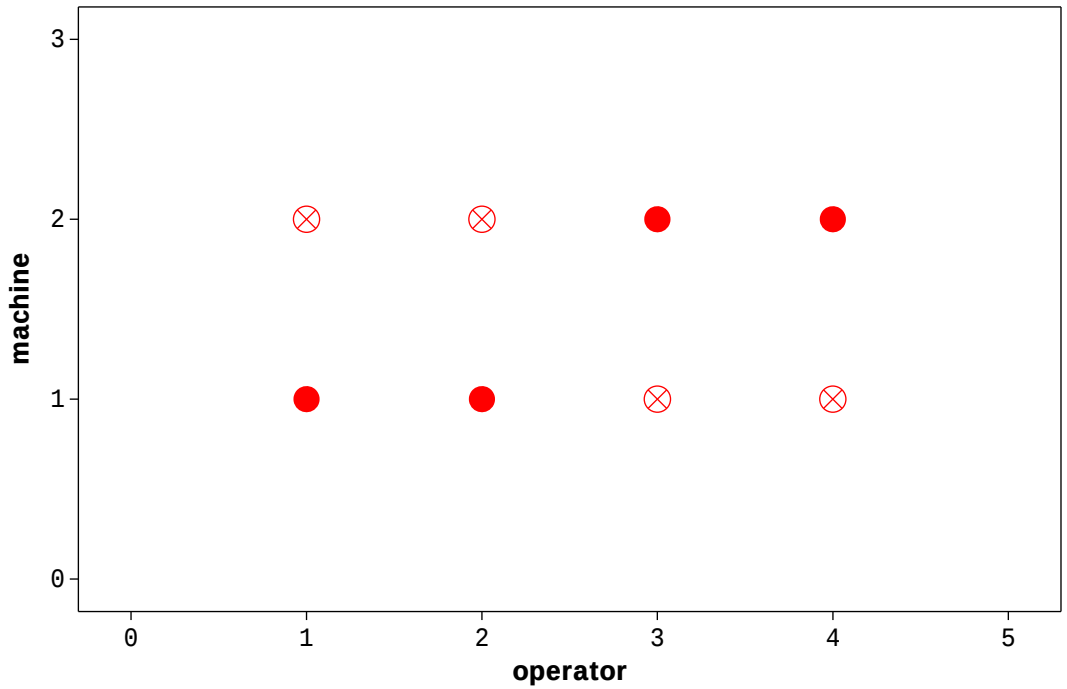


Figure: Example of two factor experiment with factor B operator nested within A machine

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

ANOVA-Model for Crossed Factors

Model

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

with the usual constraints on main effects α_i, β_j and interactions $(\alpha\beta)_{ij}$

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

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

GLM-Model for Nested Factors

Model

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

with the usual constraints on main effects α_i, $\beta_{j(i)}$, 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: health awareness study

- three states in the US participated in a health awareness study
- each state independently devised a health awareness program
- three cities within each state were selected for participation and five households within each city were randomly selected to evaluate the effectiveness of the program
- a composite index (a count number) was formed (the large the index, the greater the awareness)

the data have the following hierarchical structure:

data:

		household				
state	city	1	2	3	4	5
1	1	42	56	35	40	28
1	2	26	38	42	35	53
1	3	34	51	60	29	44
2	1	47	58	39	62	65
2	2	56	43	65	70	59
2	3	68	51	49	71	57
3	1	19	36	24	12	33
3	2	18	40	27	31	23
3	3	16	28	45	30	21

Poisson model with nested random effects

for the health awareness index Y_{ijk} for household k, in city j, and state i:

$$\log E[Y_{ijk}]_{|\alpha_i,\beta_{j(i)}} = \log \mu_{ijk} = \mu + \alpha_i + \beta_{j(i)}$$

with

- a state random effect $\alpha_i \sim N(0, \sigma_S^2)$
- a city (town) random effect $\beta_{i(j)} \sim N(0, \sigma_T^2)$ nested in the state effect
- and a Poisson error $Y_{ijk} \sim Po(\mu_{ijk})$

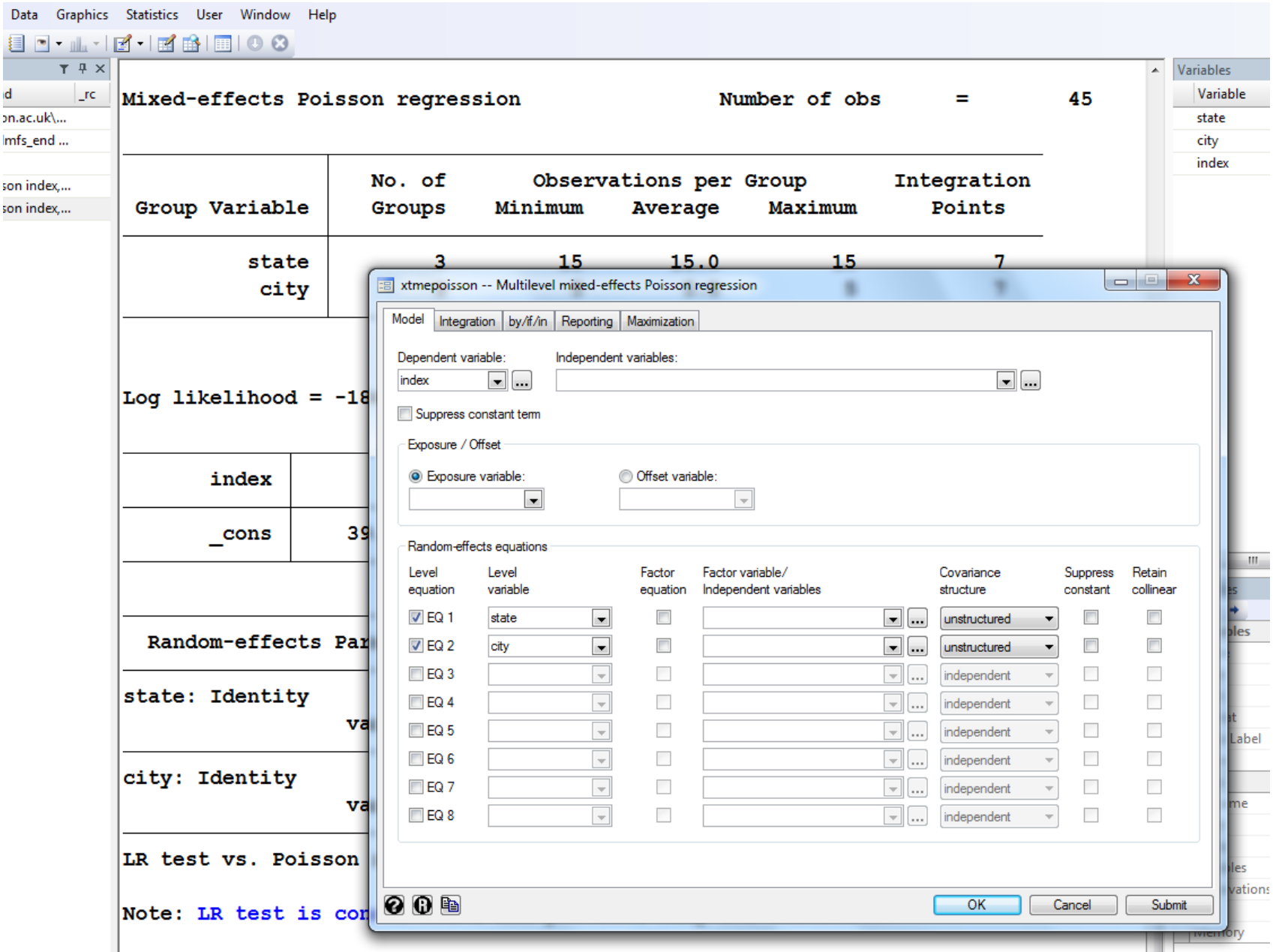
characteristics of the model

- the random effects α_i and $\beta_{j(i)}$ are usually assumed independent
- variance component model with **three variance components**: the two normal random effects of city nested in state and state itself as well as a Poisson household random error
- presence of city or state effect could be tested by

$$H_0 : \sigma_S^2 = 0 \text{ vs. } H_1 : \sigma_S^2 > 0$$

or

$$H_0 : \sigma_T^2 = 0 \text{ vs. } H_1 : \sigma_T^2 > 0$$


Data Graphics Statistics User Window Help
Mixed-effects Poisson regression Number of obs = 45
Group Variable | No. of Groups | Observations per Group Minimum | Average | Maximum | Integration Points
state | 3 | 15 | 15.0 | 15 | 7
city
Log likelihood = -18
index
_cons 39
Random-effects Par
state: Identity
va
city: Identity
va
LR test vs. Poisson
Note: LR test is con
Variables
Variable
state
city
index
xtmepoisson -- Multilevel mixed-effects Poisson regression
Model Integration by/if/in Reporting Maximization
Dependent variable:
index
Independent variables:
Suppress constant term
Exposure / Offset
Exposure variable:
Offset variable:
Random-effects equations
Level equation
Level variable
Factor equation
Factor variable/ Independent variables
Covariance structure
Suppress constant
Retain collinear
EQ 1 state unstructured
EQ 2 city unstructured
EQ 3 independent
EQ 4 independent
EQ 5 independent
EQ 6 independent
EQ 7 independent
EQ 8 independent
OK
Cancel
Submit

Mixed-effects Poisson regression Number of obs = 45

Group Variable	No. of Groups	Observations per Group Minimum	Average	Maximum	Integration Points
state	3	15	15.0	15	7
city	9	5	5.0	5	7

Wald chi2(0) = .
Log likelihood = -184.57831 Prob > chi2 = .

index	IRR	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_cons	39.8147	7.124027	20.59	0.000	28.03743	56.53905

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
state: Identity				
var(_cons)	.0942906	.0785085	.018439	.4821678
city: Identity				
var(_cons)	7.17e-21	8.18e-12	0	.

LR test vs. Poisson regression: chi2(2) = 153.09 Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.

```
Mixed-effects Poisson regression                Number of obs      =         45
Group variable: state                           Number of groups   =          3

                                                Obs per group: min =         15
                                                               avg =       15.0
                                                               max =         15

Integration points =   7                        Wald chi2(0)       =          .
Log likelihood = -184.57831                     Prob > chi2        =          .
```

index	IRR	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_cons	39.81469	7.12406	20.59	0.000	28.03739	56.53914

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
state: Identity				
var(_cons)	.0942915	.0785098	.018439	.4821776

LR test vs. Poisson regression: chibar2(01) = 153.09 Prob>=chibar2 = 0.0000