HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Modelling hierarchically structured data with MLwiN software: Introduction

Risto Lehtonen
Department of Social Research
University of Helsinki

Lecture notes, 22-23 May 2014

Outline

- Teachers
  - Prof. Antero Malin, University of Jyväskylä
  - Prof. Risto Lehtonen, UH
- Scope (Optional): 3 cu with completed practical work
- Type: Advanced studies
- Materials:
  - Course homepage

Background

- **Hierarchically structured data** are common in quantitative research in social sciences, psychology and educational sciences
- **The hierarchical structure of the data involves correlations between observations**
  - **The correlations must be accounted for in statistical analysis**
  - **WHY: For valid statistical inference**
  - **Hierarchical or multilevel models are often used for this purpose**
- In the course, basic properties of multilevel regression and ANOVA models are introduced and demonstrated with the MLwiN software
- In addition to lecture sessions, PC training sessions will be arranged for practical application of the methods

Complex data structures

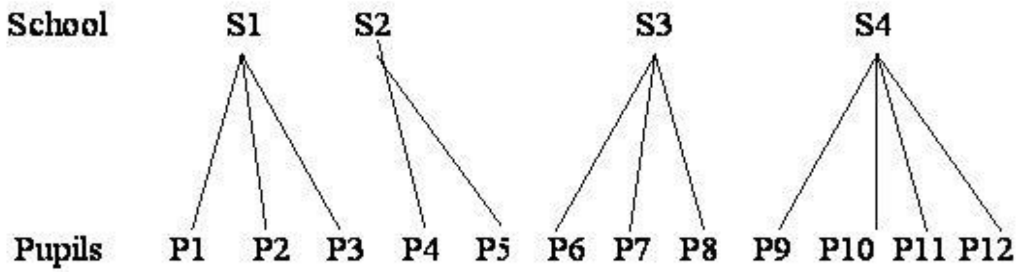
- Complex data structures are common in various areas of survey statistics
  - **Complex sampling design** involving clustering, stratification and unequal probability sampling
  - **Panel or longitudinal study design**, possibly involving rotation panels
- OECD: Programme for International Student Assessment PISA
- European Social Survey (ESS)

Clustered data structure

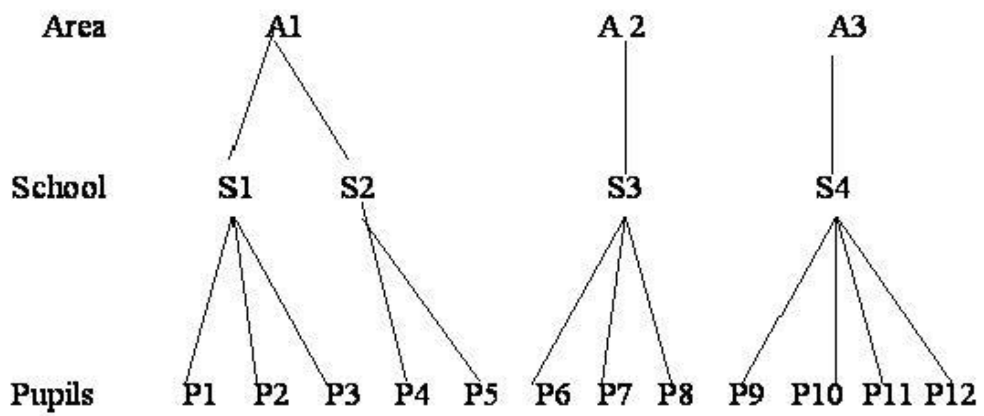
- Stratified multi-stage sampling design
- Hierarchically structured data
  Clustered data, Multilevel data
- **Cluster = a grouping containing *lower level* elements in the population or sample**
- Examples: clustered or multilevel structures
  - Schools – Students
  - Establishments – Staff members
  - Health centers – Patients
  - Neighborhoods – Households – Household members
  - Persons – measurement occasion for a person

Two-level and three-level nested structures

- Two-level nested structure with schools as clusters



- Three-level nested structure clustered by area and school



http://www.bristol.ac.uk/cmm/learning/multilevel-models/data-structures.html

Correlation of observations

- Clustered data structure involves certain type of dependence between observations called **intra-cluster correlation**
  - Cluster sampling involves **intra-cluster** (intra-class) **correlation within clusters**
  - Panel design involves **autocorrelation**
- NOTE: Elements can be assumed independent under simple random sampling SRS
  - Recall: *iid assumption = independent identically distributed random variables*
  - Corresponds SRS with replacement (SRSWR)

Hierarchical or clustered structure and sources of correlation of observations

Levels of hierarchy	Research design	
	a. Cross-sectional	**b. Longitudinal (Panel design)**
1. Single-level data (no clustering)	1a. No correlation between observations	1b. **Autocorrelation** between observations
2. Two or more levels (clustered data)	2a. **Intra-class correlation** between observations	2b. More complex covariance structures

Analysis of complex survey data

- **Key point:** Accounting for the complexities of survey data in the analysis phase ensures valid statistical inference
- Sampling design complexities
  - Multi-stage sampling design
  - Stratification and clustering
  - Weighting for unequal probability sampling
  - Weighting for unit nonresponse
  - Imputation for item nonresponse
- Study design complexities
  - Panel structure

Analysis of multilevel data

- Terminology
  - Multilevel models
  - Hierarchical models
  - Mixed models
- Linear mixed models
  - Continuous response variable
- Generalized linear mixed models GLMM
  - Continuous response – Linear mixed model
  - Binary response – Logistic mixed model
  - Polytomous response – Logistic mixed model
    - Nominal or ordinal level of measurement
  - Count response – Poisson mixed model

Generalized linear mixed model GLMM

Model:

$$E_m(y_k \mid \mathbf{u}_d) = f(\mathbf{x}'_k(\boldsymbol{\beta} + \mathbf{u}_d))$$

where $f(.)$ refers to the link function, e.g.

- linear mixed model
- logistic mixed model

$\mathbf{x}_k = (1, x_{1k}, ..., x_{pk})'$ vector of explanatory variable values for element k

$\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)'$ fixed effects

$\mathbf{u}_d = (u_{0d}, ..., u_{pd})'$ random effects

Special case 1
Linear fixed-effects model

Model:

$$E_m(y_k) = \mathbf{x}_k' \boldsymbol{\beta}$$

where

$\mathbf{x}_k = (1, x_{1k}, \ldots, x_{pk})'$ vector of explanatory variable values for element k

$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ fixed effects

E.g. $y_k = \beta_0 + \beta_1 x_{1k} + \ldots + \beta_p x_{pk} + \varepsilon_k$

Special case 2
Linear mixed model

Model:

$$E_m(y_k \mid \mathbf{u}_d) = \mathbf{x}_k'(\boldsymbol{\beta} + \mathbf{u}_d)$$

where

$\mathbf{x}_k = (1, x_{1k}, \ldots, x_{pk})'$ vector of explanatory variable values for element *k*

$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ fixed effects

$\mathbf{u}_d = (u_{0d}, \ldots, u_{pd})'$ cluster-specific random effects

E.g. $y_k = \beta_0 + u_{0d} + \beta_1 x_{1k} + \ldots + \beta_p x_{pk} + \varepsilon_k$

Special case 3
Logistic fixed-effects model

Model

$$E_m(y_k) = \frac{\exp(\mathbf{x}_k'\boldsymbol{\beta})}{1+\exp(\mathbf{x}_k'\boldsymbol{\beta})}$$

where

$\mathbf{x}_k = (1, x_{1k}, \ldots, x_{pk})'$ vector of explanatory variable values for element k

$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ fixed effects

Special case 4
Logistic mixed model

Model

$$E_m(y_k | \mathbf{u}_d) = \frac{\exp(\mathbf{x}_k'\boldsymbol{\beta} + \mathbf{u}_d)}{1 + \exp(\mathbf{x}_k'\boldsymbol{\beta} + \mathbf{u}_d)}$$

where

$\mathbf{x}_k = (1, x_{1k}, ..., x_{pk})'$ vector of explanatory variable values for element k

$\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)'$ fixed effects

$\mathbf{u}_d = (u_{0d}, ..., u_{pd})'$ cluster-specific random effects

Software for multilevel modeling

- MLWIN
  - Multilevel (generalized linear mixed) modeling
- HLM
  - Hierarchical (linear mixed) modeling
- MPLUS
  - Structural equation modeling (SEM)
- MIXED and GLIMMIX (SAS)
- GLIMMIX (SAS)
  - Generalized linear mixed modeling
- GLLAMM (Stata)
  - Generalized linear latent and mixed modeling
- LISREL
  - Structural equation modeling (SEM)

Capabilities of software: Aspects

- Coverage of model types
  - MLM - Multilevel modelling (Mixed models)
  - SEM analysis - Structural Equation Models
- Coverage of members of GLMM's
  - Continuous responses - Linear models
  - Binary responses - Binomial logistic models
  - Polytomous responses - Multinomial logistic models
  - Count data - Poisson regression models
- Accounting for research design complexities
  - Stratification
  - Clustering
  - Weighting

Capabilities of selected software 1
(adjusted from Chantala et al. 2005)

	SEM Analysis	MLM Analysis	Adjust for Clustering	Adjust for Stratification
MPLUS	Yes	Yes	Yes	Yes
LISREL	Yes	Yes	Yes	Yes
GLLAMM (Stata)	Yes	Yes	Yes	
MLWIN		**Yes**	**Yes**	
HLM		Yes	Yes	
MIXED (SAS)		Yes	Yes	
GLIMMIX (SAS)		Yes	Yes	

Capabilities of selected software 2

(adjusted from Chantala et al. 2005)

	Normal	Binary	Poisson	Multinomial Categorical	Ordered Categorical
MPLUS	Yes	Yes	Yes		
LISREL	Yes				
GLLAMM (Stata)	Yes	Yes	Yes	Yes	Yes
MLWIN	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**
HLM	Yes	Yes	Yes	Yes	Yes
MIXED (SAS)	Yes				
GLIMMIX (SAS)	Yes	Yes	Yes	Yes	Yes

Capabilities of selected software 3

(adjusted from Chantala et al. 2005)

	Allow MLM Sampling Weights	Method for Scaling MLM Sampling Weights	Responsibility for Scaling MLM Sampling Weights
MPLUS	Yes	Asparouhov (2006)	User
LISREL	Yes	Pfeffermann (1998)	User
GLLAMM (Stata)	Yes	Pfeffermann (1998)	User
MLWIN	**Yes**	Pfeffermann (1998)	User or MLWIN default
HLM	Yes	Normalize	HLM default
MIXED (SAS)	Yes	No explicit scaling	User
GLIMMIX (SAS)	Yes	No explicit scaling	User

Main literature (for this course)

- Goldstein H. (2003). Multilevel Statistical Models, 3rd Ed. London: Arnold.
- Goldstein H. (2011). Multilevel Statistical Models, 4th Ed. London: Arnold.
  - 2nd Edition - Downloadable, free 1995 version
- Lehtonen R. and Pahkinen E. (2004). *Practical Methods for Design and Analysis of Complex Surveys.* Second Edition. Chichester: Wiley. Section 9.4.

- MLwiN (www.cmm.bristol.ac.uk/MLwiN/

- LEMMA Learning Environment for Multilevel Methods and Applications (www.cmm.bristol.ac.uk/learning-training/index.shtml)

Supplemental literature (general)

- Chambers R.L. and Skinner C.J. (Eds.) (2004). *Analysis of Survey Data*. Chichester: Wiley.
- Chantala K, Suchindran C.M. and Blanchette D. (2005). Adjusting for Unequal Selection Probability in Multilevel Models: A Comparison of Software Packages. North American Stata Users' Group Meetings 2005 .
- Demidenko E. (2004). *Mixed Models. Theory and Applications*. New York: Wiley.
- Diggle P. J., Liang K.-Y. & Zeger S. L. (1994). *Analysis of Longitudinal Data.* Oxford: Oxford University Press.

Supplemental literature (weighting)

- Asparouhov T. (2006). General multi-level modeling with sampling weights. *Communications in Statistics: Theory and Methods*, 35, 3, 439-460.

- Pfeffermann D., Skinner C.J., Holmes D.J., Goldstein H. and Rasbash, J. (1998). Weighting for Unequal Selection Probabilities in Multilevel Models. *JRSS*, Series B, 60, 123-40.

- Additional materials, see: www.statmodel.com/resrchpap.shtml